Reinforcement Learning-based Artificial Pancreas Systems to Automate Treatment in Type 1 Diabetes

Chirath Yudara Hettiarachchi

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

November 2023

Except where otherwise indicated, this thesis is my own original work.

Chirath Yudara Hettiarachchi
6 November 2023

to my loving family and dear friends.

Acknowledgments

I acknowledge and celebrate the First Australians on whose traditional lands we meet, and pay my respect to the elders, past, present and emerging. I am incredibly honoured to be part of "Our Health in Our Hands (OHIOH)", a strategic initiative of the Australian National University (ANU), which aims to transform healthcare by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. OHIOH has amplified my abilities to pursue my passion for developing healthcare technologies to improve the quality of life of people and thus add value to the community. I wish to highlight the financial support extended by the ANU School of Computing and the OHIOH initiative to carry out my research. Furthermore, this work is also supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation Scheme.

I am sincerely grateful for my supervisors, Prof. Hanna Suominen, Dr Eleni Daskalaki, Prof. Christopher Nolan, Dr Nicolo Malagutti, and all my friends at the OHIOH Big Data Group, who constantly provided me with valuable advice and guidance in my research. Their support and motivation have also been precious for my personal development amidst the global challenges due to the pandemic. I would also like to thank my collaborators, Prof. David O'Neal, Dr Barbora Paldus, and Dr Dale Morrison from the Diabetes Technology Research Group, St Vincent's Hospital, University of Melbourne, Assoc./Prof. Jane Desborough, Dr Anne Parkinson, and Dr Nicola Brew-Sam from the OHIOH Health Experience Team, and the OHIOH Clinical Cohort (Diabetes) for their valuable advice and guidance in my research. Furthermore, I would like to thank Prof. Inger Mewburn, Dr Jillian Schedneck and everyone else from the ANU Researcher Development and the ANU Academic Skills Teams for their support in improving my research skills, the ANU library for providing valuable insights in conducting systematic literature reviews, and the administrative staff of the ANU College of Engineering, Computing and Cybernetics for their help in my PhD journey.

I wish to thank the taxpayers of Sri Lanka who contributed towards the free education I received from my schooling to university studies, which enabled me to achieve my full potential. Furthermore, all my previous supervisors at the University of Moratuwa, Department of Electronics & Telecommunication Engineering and Department of Computer Science and Engineering constantly guided me in my career and helped me pursue my passion for research.

Finally, my parents, wife, and brother have always encouraged me throughout my research career and supported me with their best. I want to extend them my heartfelt gratitude and love for being a constant force of strength in my life.

Abstract

Type 1 Diabetes (T1D) is a chronic disease, which impairs the glucose homeostasis of the body due to a deficiency in insulin production. People with T1D are life dependent on external insulin administration. Despite advancements in insulin treatment, current methods are still associated with a heavy cognitive burden on people with T1D due to the need for manual meal announcement, carbohydrate (CHO) estimation, and insulin dose calculation. To improve and automate treatment, Artificial Pancreas Systems (APS) have been introduced; they consist of a continuous subcutaneous glucose sensor, insulin pump, and control algorithm to estimate the insulin dose. However, the glucoregulatory system is a partially observable complex dynamical system with large inter- and intra-population variability, affected by unknown disturbances related to meals, exercise, and stress among others. This complexity, along with delays in glucose sensing and insulin action, has challenged the performance of existing control algorithms, which still require users to engage manually. Related research has advanced to exploring Machine Learning (ML) for APS and integrating additional physiological signals as Multi-input Artificial Pancreas Systems (MAPS).

This PhD thesis in Computer Science primarily presents a body of novel research on using Reinforcement Learning (RL), a class of ML algorithms to address control challenges in APS, where RL-based APS are designed and developed to automate insulin delivery by eliminating CHO estimation and meal announcement. Secondarily, the thesis explores and analyses the use of additional physiological signals for MAPS.

The glucose regulation problem was formulated as a continuing task using the average reward RL framework with new state-space, action-space, and reward function designs addressing real-world challenges associated with the problem. State-of-the-art RL algorithms were first explored to design RL-based APS. A novel algorithm named G2P2C (Glucose Control by Glucose Prediction and Planning) was designed and developed to address the shortcomings identified in the state-of-the-art algorithms. G2P2C integrated Proximal Policy Optimisation (PPO) with a model-learning phase —as an auxiliary learning task— and a planning phase to improve performance and safety. G2P2C did not require any prior announcement of upcoming meals or meal CHO content. Its performance was tested *in-silico* using an open-source T1D simulator based on the Food and Drug Administration-approved UVA/PADOVA 2008 model and benchmarked against clinical treatment strategies and the state-of-the-art RL algorithms. G2P2C achieved a time-in-range of 73% and 64% for adult and adolescent cohorts, respectively. This meant it outperformed basal-bolus clinical treatment strategies in the adult cohort without requiring any meal user input. The control performance and algorithmic characteristics of G2P2C show

promise as a candidate algorithm for automating glucose control in APS.

The potential of MAPS was first explored by conducting a systematic literature review of 17 shortlisted publications from Scopus, PubMed, and IEEE Xplore to identify and analyse input signals of MAPS. Heart rate, accelerometer readings, energy expenditure, and galvanic skin response were the main signals found that can be captured by non-invasive wearable devices, while evidence was also found for lactate and adrenaline as potential invasive biomarkers of T1D. Invasive biomarkers were explored less, while identified additional signals mainly focused on exercise in people with T1D which is challenging for existing treatment due to the rapidly changing insulin needs related to different types of exercise. Therefore, next, motivated by the review findings relationship of invasive physiological signals with exercise types (e.g., Moderate-Intensity Exercise (MIE), High-Intensity Exercise (HIE), and Resistance Exercise (RE)) that affect glucose control in T1D was investigated. A multivariate transfer entropy approach was used to model the physiological signals to learn the inter-relations of the glucoregulatory system. The analysis validated lactate as an important biomarker in estimating exercise events, while noradrenaline was observed to be valuable in differentiating RE from HIE and MIE. This highlighted the value of MAPS as the next generation of APS.

The *in-silico* evaluations of the RL-based APS show promise for eliminating meal announcement and CHO estimation in glucose regulation, and the analysis on physiological signals encourages future research into MAPS which are valuable outcomes for the T1D diabetes community. The proposed novel G2P2C algorithm and solutions to address challenges associated with using RL in real-world problems benefit the RL community. This multidisciplinary research also focused on bridging existing barriers between clinicians, lived experience experts, engineers, and researchers by publishing across broad audiences in international peer-reviewed venues and developing software tools. This kind of communication between these stakeholders is often neglected in existing research, resulting in systems that may not reflect the clinical objectives and expectations of people with T1D. The inherent complexity of ML algorithms further exacerbates the situation where the understanding and explainability of the systems, their operation, outcomes, and data analysis mechanisms are low, leading to delays and failures in adoption. Hence, as part of this thesis, an online interactive demonstration tool named CAPSML was developed to facilitate communication required to develop APS for T1D and a Graphical Processing Unit (GPU)-based T1D simulation environment named GluCoEnv was developed to encourage research and development. The codebase and software tools developed in the thesis were released open-source under the MIT license for reproducibility and to encourage future work in this domain.

Keywords: Algorithms; Closed Loop Systems; Control Systems; Diabetes Mellitus, Type 1; Glycemic Control; Insulin Infusion Systems; Machine Learning; Medical Informatics; Pancreas, Artificial; Open Source Software; Reinforcement Learning.

Publications and Software Tools

This thesis is the result of my original work during my PhD studies at the Australian National University. All references to external sources and collaborations with colleagues have been duly acknowledged where applicable. The research outcomes of this thesis have a broader societal impact due to the transdisciplinary and multi-stakeholder nature. Hence, most of the work presented in the thesis is published in computer science and clinical venues targeting the diverse academic community. In addition, two software tools were developed to reach out to the non-academic community (clinical practitioners, patients, and software developers). Furthermore, the codebase of this thesis work is released as open-source to encourage future research in this domain. The software tools and codebase are provided under the MIT license.

Software Tools and Codebase

- CAPSML: https://capsml.com/
- GluCoEnv: https://github.com/chirathyh/GluCoEnv
- Codebase: https://github.com/chirathyh/G2P2C.

Journal publications

- Hettiarachchi, C., Daskalaki, E., Desborough, J., Nolan, C.J., O'Neal, D. and Suominen, H., **Integrating Multiple Inputs Into an Artificial Pancreas System: Narrative Literature Review**, *JMIR Diabetes*, 7(1), e28861, February 2022.
- Hettiarachchi, C., Malagutti, N., Nolan, C., Suominen, H. and Daskalaki, E., **G2P2C – A Modular Reinforcement Learning Algorithm for Glucose Control by Glucose Prediction and Planning in Type 1 Diabetes**, *Biomedical Signal Processing & Control*, Elsevier, May, 2023 (Pending Review).

Conference publications

- Hettiarachchi, C., Malagutti, N., Nolan, C., Suominen, H. and Daskalaki, E., **Non-linear Continuous Action Spaces for Reinforcement Learning in Type 1 Diabetes**, *35th Australasian Joint Conference on Artificial Intelligence (AJCAI)*, Proceedings (pp. 557-570), Springer, Cham, December 2022.
- Hettiarachchi, C., Malagutti, N., Nolan, C., Daskalaki, E. and Suominen, H., **A Reinforcement Learning Based System for Blood Glucose Control without Carbohydrate Estimation in Type 1 Diabetes: In Silico Validation**, *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Proceedings (pp. 950-956), IEEE, July 2022.

Demonstration publication in international peer-reviewed proceedings

- Hettiarachchi, C., Malagutti, N., Nolan, C., Daskalaki, E., and Suominen, H., **CAPSML: Bridging the Gap Between Clinicians, Lived Experience Experts, and Artificial Intelligence Systems for Glucose Regulation in Type 1 Diabetes**, *34th Medical Informatics Europe (MIE)*, May 2023, (Accepted).

Abstracts in international peer-reviewed proceedings

- Hettiarachchi, C., Malagutti, N., Nolan, C., Suominen, H. And Daskalaki, E., **Deep Reinforcement Learning for Eliminating Carbohydrate Estimation In Glucose Regulation In Type 1 Diabetes**, *Diabetes Technology & Therapeutics*, Proceedings (Vol. 24, Pp. A95-A95), Mary Ann Liebert Inc, April 2022.

- Hettiarachchi, C., Daskalaki, E., Malagutti, N., Nolan, C. And Suominen, H., **Model-Free Inference of Information Flow Among Physiological Signals In Type 1 Diabetes Subjects Using Multivariate Transfer Entropy**, *Diabetes Technology & Therapeutics*, Proceedings (Vol. 23, Pp. A99-A99), Mary Ann Liebert Inc, June 2021.

Co-authored publications

- Weng, H., Hettiarachchi, C., Nolan, C., Suominen, H. and Lenskiy, A., **Ensuring Security of Artificial Pancreas Device System Using Homomorphic Encryption**, *Biomedical Signal Processing & Control*, 79, p.104044, Elsevier, January 2023.

- Cui, R., Hettiarachchi, C., Nolan, C.J., Daskalaki, E. and Suominen, H., **Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network**, *IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, Proceedings (pp. 154-159), IEEE, June 2021.

Notation

Below is a list of mathematical notation used in this thesis.

$\mathbb{R}$	set of real numbers
$\doteq$	equality relationship that is true by definition
$\approx$	approximately equal
$f : X \to y$	function f from elements of set X to elements of set y
$\in$	is an element of ; e.g., $x \in X$
$\ln x$	natural logarithm of x
e^x	the base of the natural logarithm, carried to power x; $e^{\ln x} = x$
$(a, b]$	the real interval between a and b including b but not including a
$Pr\{X = x\}$	probability that a random variable X takes on the value x
$X \sim p$	random variable X selected from a distribution $p(x) \doteq Pr\{X = x\}$
$\mathbb{E}[X]$	expectation of a random variable X; $\mathbb{E}[X] \doteq \sum_x p(x)x$
$H(X)$	entropy of random variable X
$\hat{x}$	estimated value of x
$\arg\max_a f(a)$	a value of a at which $f(a)$ takes its maximal value
d_{KL}	Kullback-Leiber divergence
$\leftarrow$	assignment

Application specific notation:

g_t	glucose sensor measurement at time t measured in mg/dL
i_t	insulin administered at time t measure in insulin units (U)
c_t	meal carbohydrate estimate at time t measured in grams (g)
I_{pump}	insulin infusion rate of the insulin pump (U/min)
I_{basal}	basal insulin infusion rate (U/min)
I_{max}	maximum insulin infusion rate of the insulin pump (U/min)

In Reinforcement Learning:

S	set of all states
$\mathcal{A}$	set of all actions
O	the observation function
R	the reward function
s, s'	states
a	an action
r	a reward
t	discrete time step
a_t	action at time t
$a_{t:t+n}$	sequence of actions from time t to $t + n$
s_t	state at time t

r_t	reward at time t
π	policy (decision-making rule)
$\pi(a\|s)$	probability of taking action a in state s under a *stochastic* policy π
h	horizon, the time step one looks up to in a forward view
R_t^{avg}	average reward at time t
γ	discount-rate parameter
G_t	return following time t
$v^\pi(s_t)$	value of state s_t under policy π (expected return)
$q^\pi(s_t, a_t)$	value of taking action a_t in state s_t under policy π
$A^\pi_{s_t,a_t}$	advantage of taking action a_t in state s_t under policy π
θ, ϕ	parameters of neural networks
∇_θ	the gradient of θ
π_θ	policy function parameterised by θ
v_ϕ	value function parameterised by ϕ
Q_ϕ	Q-value function parameterised by ϕ

Contents

Figures

Tables

Abbreviations

A2C	Advantage Actor Critic.
AI	Artificial Intelligence.
APS	Artificial Pancreas Systems.
ARX	Auto-Regression with Exogenous input.
BB	Basal-Bolus.
BBHE	Basal-Bolus Human Error.
BBI	Basal-Bolus Ideal.
CGM	Continuous Glucose Monitoring.
CHO	Carbohydrate.
CIR	Carbohydrate Insulin ratio.
CNN	Convolutional Neural Networks.
CPU	Central Processing Unit.
CSGM	Continuous Subcutaneous Glucose Monitoring.
CSII	Continuous Subcutaneous Insulin Infusion.
DL	Deep Learning.
DQN	Deep Q Learning.
ECG	Electrocardiogram.
EE	Energy Expenditure.
FDA	Food & Drug Administration.
G2P2C	Glucose Control by Glucose Prediction & Planning.
GluCoEnv	Glucose Control Environment.
GPC	Generalised Predictive Control.
GPU	Graphical Processing Unit.
GSR	Galvanic Skin Response.
HBGI	High Blood Glucose Index.
HCL	Hybrid Closed-Loop.
HIE	High Intensity Exercise.
HR	Heart Rate.

ISF	Insulin Sensitivity Factor.
LBGI	Low Blood Glucose Index.
LQG	Linear Quadratic Gaussian.
LSTM	Long-Short Term Memory.
MAPS	Multi-input Artificial Pancreas Systems.
MCTS	Monte Carlo Tree Search.
MDI	Multiple Daily Injections.
MDP	Markov Decision Process.
MET	Metabolic Equivalent.
MIE	Moderate Intensity Exercise.
ML	Machine Learning.
MLC	Machine Learning Control.
MPC	Model Predictive Control.
MTE	Multivariate Transfer Entropy.
PID	Proportional Integral Derivative.
POMDP	Partially Observable Markov Decision Process.
PPG	Phasic Policy Gradient.
PPO	Proximal Policy Optimisation.
RE	Resistance Exercise.
RI	Risk Index.
RL	Reinforcement Learning.
RNN	Recurrent Neural Networks.
SAC	Soft Actor Critic.
SAP	Sensor Augmented Pump.
SI	System Identification.
SMBG	Self-Monitoring of Blood Glucose.
SVM	Support Vector Machines.
T1D	Type 1 Diabetes.
TAR	Time above Range.
TBR	Time below Range.
TCN	Temporal Convolutional Networks.
TDI	Total Daily Insulin.
TIR	Time in Range.

Chapter 1

Introduction

This PhD thesis presents a transdisciplinary research focusing on computer science in healthcare. In this chapter, a high-level introduction to the research problem of glucose regulation in Type 1 Diabetes is presented, followed by an overview of machine learning and reinforcement learning methods explored in the thesis. Finally, the motivation and aims of the research are presented along with an outline of the thesis and a summary of key contributions.

1.1 The Research Problem: Glucose Regulation in Type 1 Diabetes

Type 1 Diabetes (T1D) is a chronic disease affecting millions of people worldwide, leading to a life-long optimisation problem of blood glucose regulation [DiMeglio et al., 2018]. The endocrine pancreas maintains glucose homeostasis through regulated insulin secretion in healthy individuals. However, in people with T1D, this process fails due to autoimmune destruction of the insulin-producing pancreatic islet β-cells. Hence, an appropriate amount of insulin must be administered from exogenous sources to control blood glucose concentration. Glucose control is challenging due to the varying insulin requirements related to sleep patterns, meals, and exercise. During periods of fasting (e.g., during sleep), low levels of insulin in blood are required, whereas after meals, surges in blood insulin are necessary to maintain normal blood glucose concentrations [Rorsman et al., 2000]. The effects of exercise on glucose concentrations are complicated due to the different types of exercise (e.g., moderate-intensity, high-intensity, resistance exercise) [Paldus et al., 2022b]. An inability to match insulin delivery with an individual's changing insulin requirements results in either hypoglycaemia (low blood sugar) or hyperglycaemia (high blood sugar). Hypoglycaemia, if severe, may lead to loss of consciousness, seizures or even death. Long-term exposure to hyperglycaemia results in complications such as blindness, limb amputations, kidney disease, and cardiovascular disease [DiMeglio et al., 2018]. To this end, maintaining glucose homeostasis continuously is of significant importance to avoid severe short and long-term complications.

Artificial Pancreas Systems (APS) [Cobelli et al., 2011] are a T1D treatment approach that aims to mimic the action of the pancreas by automatically releasing

insulin based on changing glucose levels in a closed-loop manner. APS are being developed to improve glucose control compared to sub-optimal open-loop treatment schemes and to remove the cognitive burden on people with T1D and their caregivers by automating insulin treatment. APS are based on a sensor/actuator mechanism, containing an external device to measure glucose levels and a pump to infuse the required insulin dose. A control algorithm calculates the required insulin dose, which closes the loop between the two devices (Figure 1.1). Although the concept of APS have been present for more than 40 years, the first Food and Drug Administration (FDA)-approved commercial system (Medtronic Minimed 670G) was only released in 2016 in the United States of America[1] [Sherwood et al., 2020]. In existing systems, the glucose levels in subcutaneous interstitial fluid are continuously monitored through a continuous glucose monitor (CGM). Then, insulin is infused subcutaneously via a pump attached to the body [Saunders et al., 2019; Breton and Kovatchev, 2021; Leelarathna et al., 2021].

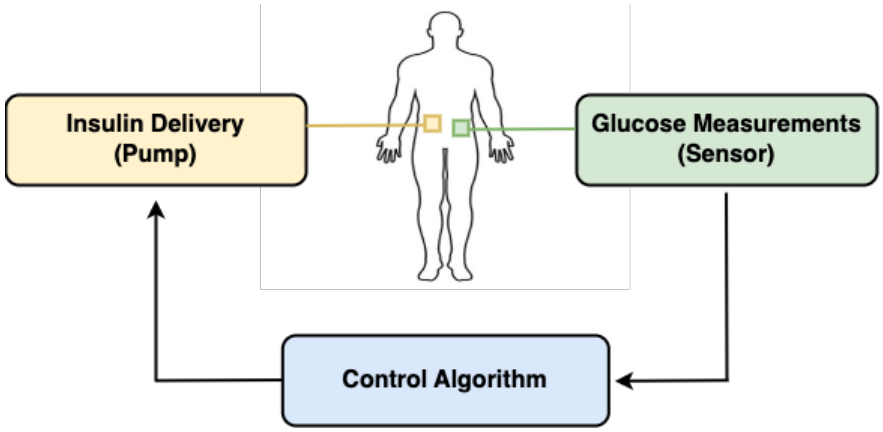


Figure 1.1: Artificial pancreas system.

APS use Proportional Integral Derivative (PID) [Steil, 2013] controllers and Model Predictive Controllers (MPC) [Bequette, 2013] to calculate the insulin requirement. However, APS are still hybrid in nature and referred to as Hybrid Closed-Loop (HCL) systems. They are not fully automatic and require manual meal announcement and calculation of insulin doses (***bolus*** insulin), a process that leads to sub-optimal glucose control due to human error [Brazeau et al., 2013; Trief et al., 2016] and adds a substantial cognitive burden on users [Brew-Sam et al., 2021a]. In particular, users must first estimate their meal's carbohydrate (CHO) content, typically 20 minutes before consumption and enter this amount into the pump system. The accuracy of each meal-related insulin bolus also depends on the insulin to CHO parameter settings entered into the pump system by the user [Diabetes Control and Complications Trial Research Group, 1993; Slattery et al., 2018]. Depending on the HCL system being used, the user may also have to decide whether to manually initiate correc-

[1]The Therapeutic Goods Administration (TGA) in Australia approved the first commercial system in 2019 [Carolyn et al., 2020].

tion insulin boli according to insulin sensitivity factor settings [Diabetes Control and Complications Trial Research Group, 1993; Slattery et al., 2018]. Hence, these systems are fully automated only for controlling the background insulin levels (also known as ***basal*** insulin), which are associated with fasting periods and underlie the meal boli during the active part of the day based on manual user input. Furthermore, daily events such as meals, exercise, sleep, and stress complicate the regulation of glucose, which, as a result, degrades the performance of current systems [Cinar, 2017]. These challenges in existing HCL systems highlight the importance of fully automated APS capable of handling complex insulin requirements to improve glucose regulation and eliminate the cognitive burden. To this end, the latest research explores the use of Machine Learning (ML) techniques [Samuel, 1959] and the integration of additional physiological signals to design Multi-input Artificial Pancreas Systems (MAPS) [Cinar, 2017].

1.2 An Overview of Machine Learning and Reinforcement Learning

Various concepts and methods are being explored to achieve Artificial Intelligence (AI) [Minsky, 1961] in which ML refers to a subset of algorithms capable of automatically learning and improving without explicit instructions. ML algorithms can assess patterns in complex datasets by analysing and drawing inferences and are used in many applications such as image recognition, speech recognition, drug discovery, and genomics, where significant improvements have been achieved [LeCun et al., 2015]. The glucoregulatory system of the human body can be described as a complex dynamical system. The recent advancements in ML [Jordan and Mitchell, 2015; LeCun et al., 2015] techniques have increased focus on data-driven approaches to model such dynamical systems and design control algorithms. The development of continuous glucose sensors and pumps in T1D management provides large amounts of valuable data which has enabled the use of ML techniques effectively in T1D applications. This highlights the appropriateness of the use of ML towards T1D management. A significant emphasis on applying ML techniques to improve diabetes management can be seen in recent years [Contreras and Vehi, 2018]. These efforts mainly relate to glucose control [Tejedor et al., 2020], biomarker identification [Kavakiotis et al., 2017], blood glucose prediction [Hameed and Kleinberg, 2020], blood glucose dynamics modelling [Woldaregay et al., 2019], detection of adverse glycaemic events [Contreras and Vehi, 2018], and personalised decision support systems [Kavakiotis et al., 2017; Woldaregay et al., 2019].

Reinforcement Learning (RL) is an emerging class of ML algorithms where an intelligent agent learns to make sequential decisions to solve a given problem [Sutton and Barto, 2018]. The agent gains experience by interacting with the underlying problem environment through decision-making and receives a reward for its behaviour based on the outcomes. The rewards are designed to reflect the desired objective of the problem. The agent's goal is to maximise the cumulative reward it can achieve

which leads to the learning of a desirable behaviour [Sutton and Barto, 2018]. In 2016, an RL agent named AlphaGo was the first to defeat the world champion of the game Go, which was considered a challenging problem for Artificial Intelligence (AI) due to the complexity [Silver et al., 2016]. Ever since, RL algorithms have been successfully applied in various challenging games, where they have demonstrated superhuman performance capability/potential [Mnih et al., 2013; Schrittwieser et al., 2020]. The latest research on RL focuses on continuous control problems such as 3D humanoid motion problems, robotics, and physics simulations [Schulman et al., 2017; Haarnoja et al., 2018], intelligent transportation systems [Haydari and Yılmaz, 2020], commercial cooling systems [Luo et al., 2022], and healthcare applications [Liu et al., 2020]. The problem of glucose regulation in T1D requires sequential insulin dosing decisions [DiMeglio et al., 2018], and due to the potential of RL algorithms to perform well in stochastic environments under unknown delays, in learning complex dynamics, and personalisation, previous work has explored the development of RL-based APS [Bothe et al., 2013]. Real-world RL systems are presented with many challenges, such as partial observability, high-dimension state/action spaces, large sensor/actuator delays, safety constraints, formulation of reward functions, explainability, and practical learning constraints [Dulac-Arnold et al., 2019], which are also present in the problem of glucose regulation. Hence, glucose control can be identified as a highly suitable application for developing real-world RL algorithms.

1.3 Motivation and Aim

Recent studies have identified that APS technologies have often been misunderstood, over-promised, and under-delivered [Kesavadev et al., 2020]. This calls for the co-creation across disciplines, where people with T1D and their caregivers, clinicians, health experience experts, and researchers work actively towards the design and development of APS [Desborough et al., 2022; Brew-Sam et al., 2023]. A survey conducted by Brew-Sam et al. [2021b] targeting young people with T1D and their parents regarding the experience of existing diabetes technologies further substantiates this approach;

> *"I feel like you reach a point where we kind of know a bit more [than the doctor]... because we're the ones experiencing it kind of every day." [Brew-Sam et al., 2021b, p.12].*

Existing APS technologies are HCL systems, which require further improvements to regulate glucose better and alleviate the cognitive burden on people with T1D and their caregivers [Brew-Sam et al., 2021a]. Fully automated APS can eliminate the burden on the user and avoid sub-optimal glucose control resulting from human error associated with existing systems [Brazeau et al., 2013; Trief et al., 2016].

This PhD thesis in Computer Science explores the development of fully automated APS. To this end, the primary focus is to design RL-based APS, while the secondary

focus is on MAPS.

RL-based APS. The potential of RL algorithms to perform well in stochastic environments under unknown delays, in learning complex dynamics, and personalisation has brought attention towards RL-based APS [Bothe et al., 2013]. However, APS are classified as high-risk medical devices, and the application of RL to real-world problems is presented with numerous research, engineering, and practical challenges along with critical ethical considerations [Dulac-Arnold et al., 2019]. Hence, the development of APS and the application of RL to design APS remain an open challenge. Therefore, this thesis first directs attention to exploring a suitable RL-based problem formulation for fully automated glucose control, considering the challenges of applying RL to the real world. Next, the state-of-the-art RL algorithms are used to design RL-based APS and evaluated based on their performance and algorithmic characteristics. Finally, the thesis aims to develop a novel modular RL algorithm to overcome limitations in the state-of-the-art RL algorithms, specifically focusing on safety and performance improvements.

MAPS. The rapid development of wearable technologies has presented access to additional information by capturing physiological signals such as heart rate, energy expenditure, and galvanic skin response [Iqbal et al., 2016]. In addition, invasive sensors capable of capturing blood biomarkers such as lactate and adrenaline are also currently being developed. These recent advancements present an opportunity to utilise the additional information to assist glucose control by detecting daily events such as exercise and stress and addressing the uncertainty and disturbances which affect existing APS. Hence, integrating additional signals to APS to design MAPS is a promising area of research [Cinar, 2017]. Therefore, this thesis conducts a systematic literature review to explore the effects of identified additional signals. Next, based on the review findings, relatively less explored invasive physiological signals are analysed in the context of different exercise types in people with T1D to understand their effects on the glucoregulatory system.

In order to better understand the user preferences and application requirements of T1D and better communicate the research of this thesis, a multidisciplinary research approach was pursued where people with T1D, clinicians, and health experience experts were consulted throughout the research. Furthermore, software tools were explored to improve the explainability of proposed systems to non-technical audiences and address engineering challenges associated with the experimental pipeline of RL-based APS. The research objectives of this thesis can be summarised as follows:

1. Develop fully automated RL-based APS to alleviate the cognitive burden on people with T1D.
2. Explore the use of additional physiological variables to develop MAPS.
3. Develop software applications/tools to support T1D education and research.

1.4 Thesis Outline and Key Contributions

In this section, the structure of the thesis is presented by summarising the content of the chapters and underlying contributions. The thesis aims to develop fully automated APS. The background and related work are presented in Chapters 2 and 3, where Chapter 2 focuses on T1D, APS, and ML, while Chapter 3 focuses on RL. To achieve the thesis objectives, the primary focus is on RL-based APS, which is presented in Chapters 4, 5, and 6. The secondary focus on MAPS is presented in Chapter 7. Chapter 8 introduces two software tools developed during this thesis to facilitate the development of APS, improve communication between the stakeholders of this multidisciplinary research, and discuss the future of RL-based APS. Finally, Chapter 9 concludes the thesis by summarising the research contributions, limitations, and the research impact.

Chapter 1. Introduction

In this chapter, a high-level introduction to the research problem of glucose regulation in T1D is presented, followed by an overview of ML and RL methods explored in the thesis. Finally, the motivation and aims of the research are presented along with an outline of the thesis and a summary of its key contributions.

Chapter 2. Background & Related Work I: Type 1 Diabetes, Artificial Pancreas Systems, and Machine Learning

This chapter summarises the preliminary concepts and background related to the thesis. First, an introduction to T1D and glucose regulation is provided, and the evolution of treatment methods leading to APS is explored. Next, APS are introduced, while highlighting the limitations of existing commercial systems and classical control research towards fully automating APS. Finally, the complexities associated with glucose regulation are summarised, and ML is introduced as a technology to overcome identified challenges.

Chapter 3. Background & Related Work II: Reinforcement Learning

This chapter provides essential background material required to develop RL-based APS. First, RL is introduced along with its concepts and applications. Next, challenges associated with applying RL to real-world problems are presented. Finally, a literature review on the use of RL for continuous control problems and APS is presented, and limitations of existing RL-based APS are highlighted to identify areas of improvement.

Chapter 4. Problem Formulation and State-Of-The-Art RL for Glucose Control

This chapter presents a fully automated RL-based APS design, targeting the cognitive burden associated with existing treatment related to CHO estimation and meal announcement. First, a formulation for the glucose control problem is proposed by focusing on the environment, state/action spaces, and reward function design. Next, the state-of-the-art RL algorithms are explored and implemented to compare against standard clinical treatment based on basal-bolus treatment strategies. Finally, the experimental design and evaluation metrics are presented, and the performance of the designed fully-automated RL-based APS are evaluated.

The work presented in this chapter is based on

- Hettiarachchi, C., Malagutti, N., Nolan, C., Daskalaki, E. and Suominen, H., **A Reinforcement Learning Based System for Blood Glucose Control without Carbohydrate Estimation in Type 1 Diabetes: In Silico Validation**, *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Proceedings (pp. 950-956), IEEE, July 2022.
- Hettiarachchi, C., Malagutti, N., Nolan, C., Suominen, H. And Daskalaki, E., **Deep Reinforcement Learning for Eliminating Carbohydrate Estimation In Glucose Regulation In Type 1 Diabetes**, *Diabetes Technology & Therapeutics*, Proceedings (Vol. 24, Pp. A95-A95), Mary Ann Liebert Inc, April 2022.

Chapter 5. G2P2C: A Novel Modular RL Algorithm for Glucose Control

This chapter introduces a novel RL-based algorithm to address the shortcomings of the state-of-the-art algorithms explored in the previous chapter. The proposed algorithm extends the Proximal Policy Optimisation (PPO) algorithm by introducing two novel optimisation phases, model learning and planning. In view of these novel added features, the algorithm was named G2P2C — Glucose Control by Glucose Prediction and Planning and is designed to focus on the safety, explainability, and translation of the research to real-world. This chapter presents the algorithm design and implementation details of G2P2C and evaluates its performance against clinical treatment benchmarks and the state-of-the-art RL algorithms.

The work presented in this chapter is based on

- Hettiarachchi, C., Malagutti, N., Nolan, C., Suominen, H. and Daskalaki, E., **G2P2C – A Modular Reinforcement Learning Algorithm for Glucose Control by Glucose Prediction and Planning in Type 1 Diabetes**, *Biomedical Signal Processing & Control*, Elsevier, May, 2023 (Pending Review).

Chapter 6. Non-linear Continuous Action Spaces for RL in Glucose Control

Continuous action spaces are a challenge faced by RL algorithms when applied to real-world problems (Chapter 3). In the case of glucose regulation, a large continuous action space is required to reflect the dynamic insulin needs. The standard approach in RL is to formulate a linear action space for the agent. Therefore, this chapter introduces non-linear continuous action spaces to overcome the challenge associated with efficiently exploring the dynamic range of insulin in glucose control. Three non-linear translation functions are proposed to map the RL action to the insulin infusion rate, inspired by the basal-bolus pattern of clinical insulin treatment practice. The proposed non-linear action spaces are compared and evaluated against the standard linear action space used in practice for RL algorithms. An action space designed through this proposed novel approach is used in the RL-based APS presented in Chapters 4 and 5.

The work presented in this chapter is based on

- Hettiarachchi, C., Malagutti, N., Nolan, C., Suominen, H. and Daskalaki, E., **Non-linear Continuous Action Spaces for Reinforcement Learning in Type 1 Diabetes**, *35th Australasian Joint Conference on Artificial Intelligence (AJCAI)*, Proceedings (pp. 557-570), Springer, Cham, December 2022.

Chapter 7. Multi-input Artificial Pancreas Systems (MAPS)

In previous chapters, RL was explored to address technical and application-specific challenges related to glucose control. This chapter explores the potential of integrating additional physiological signals to address identified challenges and design MAPS. The chapter is divided into two sections. The first section presents a systematic literature review to identify additional physiological signals suitable for glucose control. Based on the review findings, the following section analyses invasive physiological signals in the context of exercise. Exercise is a challenging activity affecting glucose control due to the rapidly changing insulin needs. Relationships between invasive physiological signals and different types of exercise are explored to assist glucose control and design MAPS.

The work presented in this chapter is based on

- Hettiarachchi, C., Daskalaki, E., Desborough, J., Nolan, C.J., O'Neal, D. and Suominen, H., **Integrating Multiple Inputs Into an Artificial Pancreas System: Narrative Literature Review**, *JMIR Diabetes*, 7(1), e28861, February 2022.
- Hettiarachchi, C., Daskalaki, E., Malagutti, N., Nolan, C. And Suominen, H., **Model-Free Inference of Information Flow Among Physiological Signals In Type 1 Diabetes Subjects Using Multivariate Transfer Entropy**, *Diabetes Technology & Therapeutics*, Proceedings (Vol. 23, Pp. A99-A99), Mary Ann Liebert Inc, June 2021.

Chapter 8. Future of Artificial Pancreas Systems Development

This chapter discusses essential characteristics of APS development by first discussing the importance of collaboration between various stakeholders associated with the research. Next, two software tools named CAPSML and GluCoEnv are introduced to improve the communication barrier and research and development processes in RL-based APS. Finally, the future of APS are discussed by highlighting important characteristics and processes related to safety, privacy, ethical considerations, and clinical trials, which need to be further explored.

The work presented in this chapter is based on

- Hettiarachchi, C., Malagutti, N., Nolan, C., Daskalaki, E., and Suominen, H., **CAPSML: Bridging the Gap Between Clinicians, Lived Experience Experts, and Artificial Intelligence Systems for Glucose Regulation in Type 1 Diabetes**, *34th Medical Informatics Europe (MIE)*, May 2023, (Accepted).
- Weng, H., Hettiarachchi, C., Nolan, C., Suominen, H. and Lenskiy, A., **Ensuring Security of Artificial Pancreas Device System Using Homomorphic Encryption**, *Biomedical Signal Processing & Control*, 79, p.104044, Elsevier, January 2023.
- Cui, R., Hettiarachchi, C., Nolan, C.J., Daskalaki, E. and Suominen, H., **Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network**, *IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, Proceedings (pp. 154-159), IEEE, June 2021.

Chapter 9. Thesis Conclusion

This chapter concludes the thesis by summarising the research contributions, limitations, and the research impact on diabetes and ML communities. The chapter highlights the developed novel RL algorithm named G2P2C, which eliminates meal announcement and CHO estimation to minimise the cognitive burden on people with T1D and the novel approach of non-linear action spaces to formulate the research problem addressing real-world challenges of RL. This thesis was limited to an *in-silico* study which used FDA-recognised simulators following best practices. The chapter highlights these limitations and their potential impact. It concludes by highlighting the engineering contributions through designed software tools and open-source code contributions to replicate this work and facilitate the future development of APS.

Chapter 2

Background and Related Work I: Type 1 Diabetes, Artificial Pancreas Systems, and Machine Learning

This chapter summarises the preliminary concepts and background related to the thesis. First, an introduction to T1D and glucose regulation is provided, and the evolution of treatment methods leading to APS is explored. Next, APS are introduced by highlighting the limitations of existing commercial systems and classical control research towards fully automating APS. Finally, the complexities associated with glucose regulation are summarised, and ML is introduced as a technology to overcome identified challenges.

2.1 Type 1 Diabetes (T1D)

The pancreas is a vital organ in the human body's digestive system, which mainly focuses on producing enzymes to digest food and hormones for blood glucose regulation. *Diabetes* is a metabolic disease caused by the deficiency of the hormone insulin produced by the islet β-cells of the pancreas. There are three main types of diabetes, namely type 1 diabetes, type 2 diabetes and gestational diabetes. Type 2 diabetes occurs due to dysfunction of the β-cell and insulin resistance, and gestational diabetes develops in women during pregnancy. T1D is caused by the autoimmune destruction of the β-cells of the pancreas. It results in complete insulin deficiency, rendering the need for external insulin administration to maintain glucose homeostasis vital to survival of people living with the disease [DiMeglio et al., 2018].

In 2021, it was estimated that 8.4 million people worldwide were affected by T1D [Gregory et al., 2022]. T1D causes a significant cognitive, and economic burden as people with T1D require daily insulin treatment, regular blood glucose monitoring, structured diabetes education, and expert medical care [Gregory et al., 2022]. In many low-income countries, access to these resources is limited, which leads to severe disability and early death. This is evident from the estimated remaining life expectancy of a 10-year-old diagnosed with T1D. In low-income countries, the mean

remaining life expectancy was 13 years, while it was 65 years for high-income countries in 2021 [Gregory et al., 2022]. It is estimated that 20% (1.8 million) of people with T1D are from low-income and lower-middle-income countries [Gregory et al., 2022]. In 2040, the number of prevalent cases is expected to increase to 13.5 - 17.4 million, which highlights the impact of T1D and its widespread nature [Gregory et al., 2022].

T1D requires external insulin administration for survival. Various treatment methods exist to estimate and administer insulin, where maintaining an optimal glucose level is the main target (Figure 2.1). The inability to maintain normal glucose levels results in either ***hypoglycaemia*** (low blood sugar) or ***hyperglycaemia*** (high blood sugar). Hypoglycaemia is very dangerous, resulting in seizures, loss of consciousness or even death. Long-term hyperglycaemia leads to complications such as blindness, renal disease, limb amputations and cardiovascular disease [DiMeglio et al., 2018]. To this end, maintaining glucose homeostasis continuously is of significant importance for avoiding severe short- and long-term complications. The glucose range of 70–180 mg/dL (3.9-10 mmol/L) has been identified as the optimal region for glucose for people with T1D, and the percentage of time spent in this range is named **Time-In-Range (TIR)**, a key metric which treatment aims to improve [Diabetes Control and Complications Trial Research Group, 1993]. In addition to TIR, there are several other metrics used to evaluate the performance of T1D treatment, which focus on the time spent in hypoglycaemia and hyperglycaemia ranges and their associated risk. These metrics were standardised (Table 2.1) by the Advanced Technologies & Treatment for Diabetes (ATTD) Congress in 2019 through consensus received from a panel of international physicians, researchers, and individuals with T1D, to address the lack of consistency in the practical application of the metrics [Battelino et al., 2019].

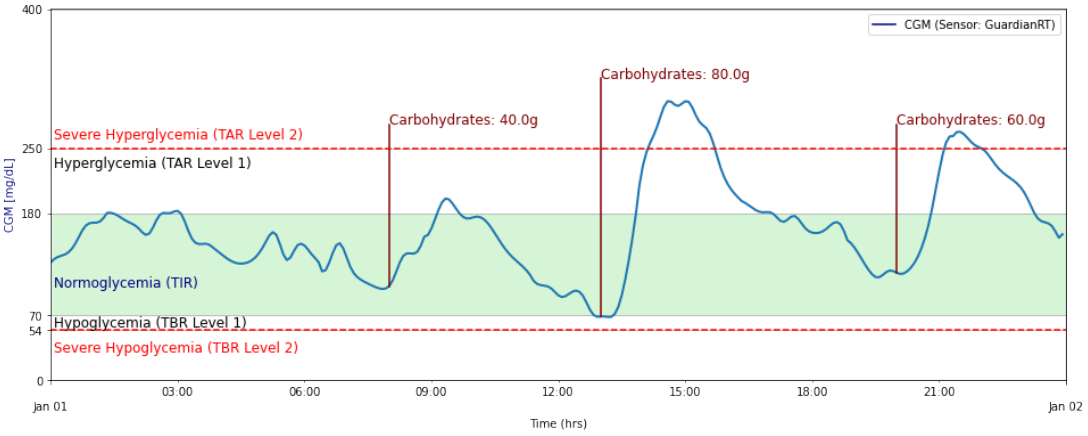


Figure 2.1: A daily simulation of glucose with meal events for type 1 diabetes.

Table 2.1: Standardised type 1 diabetes metrics.

Metric
Time above Range (TAR) - Level 2 (Severe Hyperglycaemia): Percentage time >250 mg/dL (>13.9 mmol/L)
Time above Range (TAR) - Level 1 (Hyperglycaemia): Percentage time >180 mg/dL (>10.0 mmol/L)
Time in Range (TIR) (Normoglycaemia): Percentage time 70–180 mg/dL (3.9-10.0 mmol/L)
Time below Range (TBR) - Level 1 (Hypoglycaemia): Percentage time <70 mg/dL (<3.9 mmol/L)
Time below Range (TBR) - Level 2 (Severe Hypoglycaemia): Percentage time <54 mg/dL (<3.0 mmol/L)
Low Blood Glucose Index (LBGI)
High Blood Glucose Index (HBGI)
Risk Index (RI)

2.2 Evolution of Type 1 Diabetes Treatment Methods

The discovery of insulin in 1921 led to exogenous insulin administration as a therapeutic agent, which has been life-saving for people living with T1D. Pancreas and islet cell transplants have also provided a solution for T1D more recently, though organ donation shortage, the risks of surgery, and the need for immunosuppression are limiting factors [Roep, 2020]. As a result, there is a continued reliance on the subcutaneous administration of exogenous insulin to treat T1D. The insulin delivery is divided into two distinct administration patterns: low-range, an almost-continuous basal pattern of delivery, and a pattern of intermittent high-range (or bolus) delivery of insulin used mainly to counter the glucose elevation due to meals [Nathan et al., 1993]. There have been continuous advancements in insulin preparations, including insulin analogues with different pharmacokinetic properties [Vliebergh et al., 2021]. The development of new insulin delivery and blood glucose measurement technologies have also driven improvements in T1D treatment methods (Table 2.2). Shah et al. [2016] presents the evolution of insulin delivery methods and different routes of administration, while Villena Gonzales et al. [2019] reviews the progress of glucose monitoring.

Until recent years, best practice treatment of T1D, as was established in the Diabetes Control and Complications Trial [Diabetes Control and Complications Trial Research Group, 1993], involved frequent **Self-Monitoring of Blood Glucose (SMBG)** through using finger-pricks to access capillary blood and **Multiple Daily Injections (MDI)** of short- and long-acting insulins. Information from the SMBG, as well as carbohydrate content of meals and planned exercise, informed titration of insulin doses. The development of sensors for **Continuous Subcutaneous Glucose Monitoring (CSGM)** and insulin pumps for **Continuous Subcutaneous Insulin Infusion**

Table 2.2: A summary of type 1 diabetes treatment methods.

Method	Manual Effort Required	Treatment Mechanism
MDI	Yes	Clinical guidelines
SAP	Yes	Clinical guidelines
HCL	Yes	Clinical guidelines, classical control methods (Table 2.3)
APS	No	Classical control (Table 2.4), RL (Table 3.1) methods

Acronyms: APS - Artificial Pancreas Systems, HCL - Hybrid Closed Loop, MDI - Multiple Daily Injections, RL - Reinforcement Learning, SAP - Sensor Augmented Pump.

(CSII) has improved T1D treatment. The CSGM and CSII have typically been used in an open-loop setting, in which the operator (health care professional or person with diabetes) needs to program a variable 24-hour basal infusion rate, as well as an insulin to CHO ratio (required for calculating meal insulin boluses), and an insulin sensitivity factor and target glucose range (required for calculating correction insulin boluses). However, this open-loop system is cognitively demanding because of the constant requirement for input from the patient (and carers). The CHO content of each meal and blood glucose concentrations need to be manually entered into the pump computer before initiating insulin bolus doses. In addition, basal infusion rates must be adjusted for exercise and during intercurrent illnesses. The electronic connectivity between glucose monitoring and insulin pumps has led to the recent advancement of **Sensor Augmented Pump (SAP)** therapy, which allows for automatic suspension of insulin delivery 'before low' or 'on low' to reduce the risk of severe hypoglycaemia. SAP has partially relieved the cognitive strain on people with diabetes.

Treatment of T1D based on MDI and SAP therapies uses a rule-based mechanism designed by clinicians for insulin infusion. This is carried out by calculating patient-specific characteristics such as Total Daily Insulin (TDI), Carbohydrate Insulin Ratio (CIR), and Insulin Sensitivity Factor (ISF), which are used to derive the basal insulin rate, correction insulin boli, and meal insulin boli[1] [Nathan et al., 1993; Slattery et al., 2018]. These parameters are closely monitored by clinicians and adjusted periodically. The basal insulin rate (Equation (2.1)) is associated with fasting periods (e.g., sleeping and prior to eating). The correction boli (Equation (2.2)) are used to counter the increase in measured glucose (G_m) from the target glucose level (G_t). The meal boli are used to target the increase in glucose due to meals, which is calculated by estimating the CHO content of the meal in grams (Equation (2.3)). An example glucose simulation based on SAP therapy is presented in Figure 2.2, where a fixed basal insulin rate is applied continuously, and bolus insulin doses are to counter meal events.

[1]Different basal infusion rates, CIRs, and ISFs can be set according to the time of day (e.g. morning, afternoon, evening or over night).

$$\text{Basal Insulin Rate} = 0.48 \cdot \text{TDI U/day}. \tag{2.1}$$

$$\text{Correction Insulin Bolus} = \frac{(G_m - G_t)}{\text{ISF}} \text{ U}. \tag{2.2}$$

$$\text{Meal Insulin Bolus} = \frac{\text{CHO}}{\text{CIR}} \text{ U}. \tag{2.3}$$

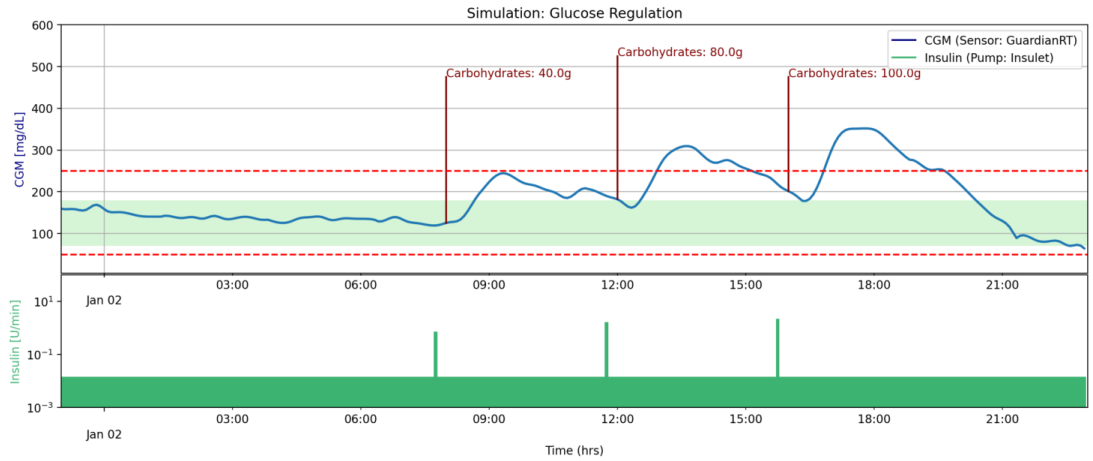


Figure 2.2: A simulation of glucose control based on sensor augmented pump therapy. Insulin (bottom) is administered to control glucose levels (top). The spikes in insulin represent the bolus insulin, while the rest represent the basal insulin. Daily meal events and their carbohydrate contents are presented in red.

2.3 Artificial Pancreas Systems (APS)

The advent of rapid-acting insulin analogues, CSGM, and CSII, have made progress towards APS possible. The essential components of an APS are a sensor measuring subcutaneous interstitial fluid glucose on a near-continuous basis, a pump infusing rapid-acting insulin into the subcutaneous tissue and a control algorithm that closes the loop between the two devices (Figure 2.3). While the concept of APS have been around for over 40 years, with the Biostator (1977) identified as the first closed-loop glucose controlling system or APS [Cobelli et al., 2011], it has only been the last few years that the use of APS have become a clinical reality. The controller is a vital component of APS, responsible for calculating the optimal amount of insulin required to maintain glucose homeostasis. The controller uses subcutaneous glucose measurements as the primary input to calculate and operate the required insulin infusion rate as the output.

The FDA has categorised APS as class III medical devices which involve automated drug delivery and are considered high-risk devices. Hence an Investigation

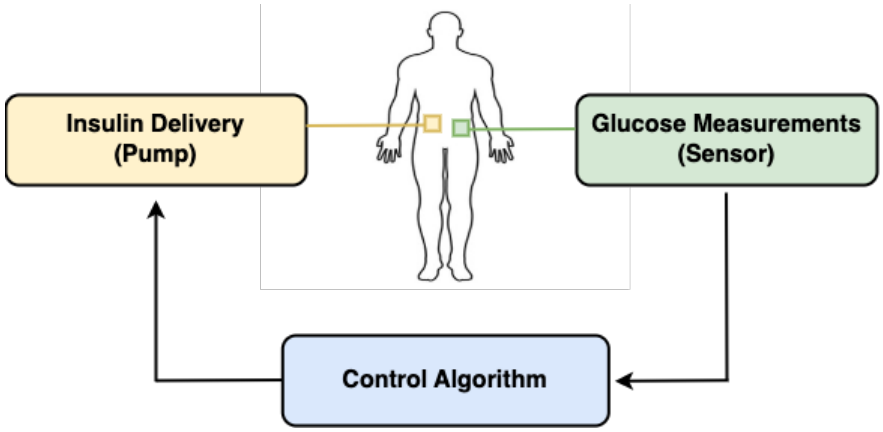


Figure 2.3: Artificial pancreas system

Device Exemption is required prior to any clinical trial [US Food and Drug Administration, 2012]. This requires initial testing of the proof of concept through animal trials, which is time-consuming and costly [Kovatchev et al., 2009]. A critical step towards APS advancement was the development of physiological models and simulators, which enabled the tuning and testing of different control algorithms *in-silico* prior to clinical studies ensuring safety. The simulators provide the capability to extensively test the robustness of control algorithms and large-scale simulations accounting for inter-subject variability compared to animal trials. The UVA/PADOVA simulator [Kovatchev et al., 2009] was accepted by the FDA as a replacement for animal testing simplifying the process, and is currently the only such approved simulator. The UVA/PADOVA simulator has been widely used in pre-clinical studies to validate proposed control algorithms.

2.3.1 Hybrid Closed-Loop Systems

The first commercial APS were released in 2016 in the United States of America[2] [Sherwood et al., 2020], with FDA-approval. Due to control algorithmic limitations to address the delay present in CSII and CSGM, these approved APS are not fully closed-loop systems but are **Hybrid Closed-Loop (HCL)** systems. Essentially, the HCL system uses closed-loop for basal insulin delivery but requires the operating person to initiate insulin boli before meals according to the CHO content and, except for the latest HCL systems, correcting boluses to counter high blood glucose levels[3]. Hence, it is a feedback system with a feedforward loop. The requirement of meal announcement and CHO estimation in HCL systems leads to a substantial cognitive burden on the users [Brew-Sam et al., 2021a]. Hence, fully closed-loop APS would

[2]The Therapeutic Goods Administration (TGA) in Australia approved the first commercial system in 2019 [Carolyn et al., 2020].

[3]The most recent commercial HCL systems provides automatic correction boluses and considered as Advanced HCL systems (AHCL) (e.g., Medtronic Minimed 780G(2020), CamAPS(2020), and Control IQ(2019)). Howver, they still require meal announcement and CHO estimation.

significantly reduce the cognitive burden on people with diabetes and their carers.

Medtronic MiniMed 670G (2016) [Messer et al., 2018], Control IQ (2019) [Forlenza et al., 2019], CamAPS (2020) [Leelarathna et al., 2020], and Medtronic MiniMed 780G (2020) [Silva et al., 2022] are currently available commercial HCL systems. All these proposed systems have conducted long-term clinical trials (3-6 months) and have been able to achieve a TIR performance of nearly 70%. Systematic reviews and meta-analysis have verified that HCL systems have shown better performance compared to conventional pump therapy [Weisman et al., 2017; Bekiari et al., 2018; Petrovski et al., 2021]. However, there is still significant room for improvement in the TIR performance. Furthermore, human errors in CHO estimation and forgetting to inform the HCL system of meals reduce the performance and increase safety concerns.

Table 2.3: Performance of commercial hybrid closed-loop systems.

System	**TIR**
Medtronic Minimed 670G (2016) [Messer et al., 2018]	69.0%
Control IQ (2019) [Forlenza et al., 2019]	71.0% ± 6.6%
CamAPS (2020) [Leelarathna et al., 2020]	67.0% ± 10.0%
Medtronic Minimed 780G (2020) [Silva et al., 2022]	76.2%± 9.1%

Acronyms: TIR - Time In Range.

2.3.2 Classical Control Algorithms

A variety of classical control algorithms have been explored in previous research towards APS development. The majority of research has extensively explored PID [Steil et al., 2006], MPC [Bequette, 2013] algorithms and their variants. PID was selected as a suitable algorithm due to the argument that it can emulate the multi-phase response of the β-cell in glucose regulation to design a more physiological insulin delivery [Steil et al., 2011]. Equation (2.4) presents a general PID algorithm, where I_{basal} is the basal insulin rate, G_m the measure of glucose, G_t the target glucose level, and K_c, τ_I, τ_D are tuneable parameters.

$$I_{PID} = I_{\text{basal}} + K_c \left[(G_m - G_t) + \frac{1}{\tau_I} \int_0^t (G_m - G_t) dt + \tau_D \frac{d(G_m - G_t)}{dt} \right]. \tag{2.4}$$

In MPC, a model of the system is used, which predicts the responses for a prediction horizon (P) based on calculated control action (u) for a given control horizon (M). The optimal control trajectory is identified by solving a constrained optimisation problem (Equation (2.5)). The optimisation objective is to minimise $J_{MPC}(\Delta u)$, which aims to control glucose levels to a target (r_{k+i}) weighted by W_y, while minimising the control input weighted by $W_{\Delta u}$. The constraints on the control action and glucose levels (y) focus on safety. Once the optimal trajectory is identified, the first control action step is applied, and the optimisation continues.

$$
\begin{aligned}
J_{MPC}(\Delta u) = & \sum_{i=1}^{P} W_y (r_{k+i} - \hat{y}_{k+i|k})^2 + \sum_{i=0}^{M-1} W_{\Delta u} \Delta u_{k+i|k}^2 \\
\text{s.t.} \quad & \Delta u_{min} \leq \Delta u \leq \Delta u_{max}, \\
& u_{min} \leq u \leq u_{max}, \\
& y_{min} \leq y \leq y_{max}.
\end{aligned} \tag{2.5}
$$

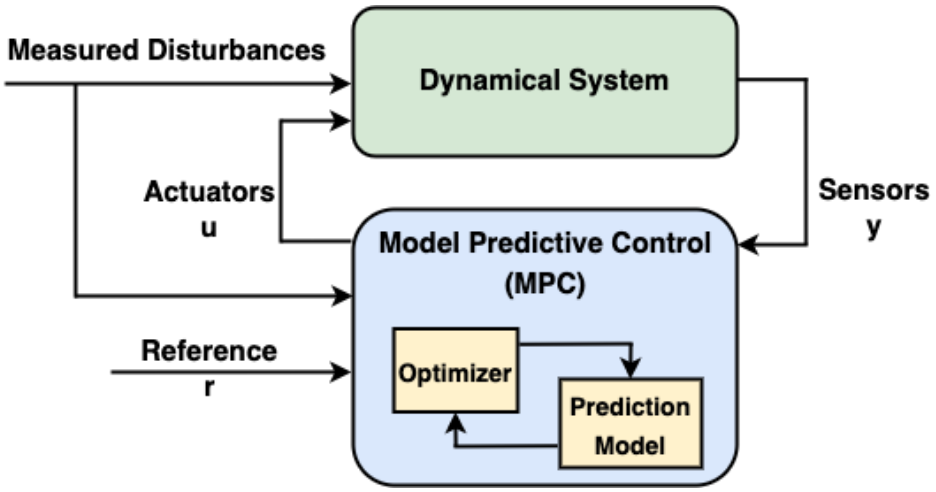


Figure 2.4: Model predictive control

Extensive comparisons between the use of PID and MPC for glucose regulation have been conducted [Bequette, 2013; Steil, 2013]. PID and MPC techniques in HCL systems require manual bolus insulin to counter disturbances from daily events such as meals and exercise. In particular, pure PID-based systems are slow in reacting to such events and, coupled with the insulin actuation delays, could lead to insulin overdose [Bothe et al., 2013]. However, PID systems are equipped with different safety mechanisms to avoid these adverse effects [Steil et al., 2011]. On the other hand, MPC approaches are more proactive as they are based on an internal system model. Generally, the underlying model used in MPC algorithms is fixed and linear and does not capture the dynamics of the disturbances (e.g., exercise). Hence, they are also prone to disturbances. Therefore, a good personalised model encompassing valuable dynamics is vital to achieving good performance. Adaptive control strategies such as Generalised Predictive Control (GPC), capable of adaptively improving the internal model to better represent the system and carry out control (Figure 2.5), have been explored for APS development [Turksoy et al., 2013d].

Linear Quadratic Gaussian (LQG) methods have also been explored in previous work [Colmegna et al., 2018]. In LQG, the full-state estimate from Kalman filtering is generally used with the full-state feedback control law from the linear quadratic regulator (Figure 2.6). In contrast to methods that rely on dynamical representations,

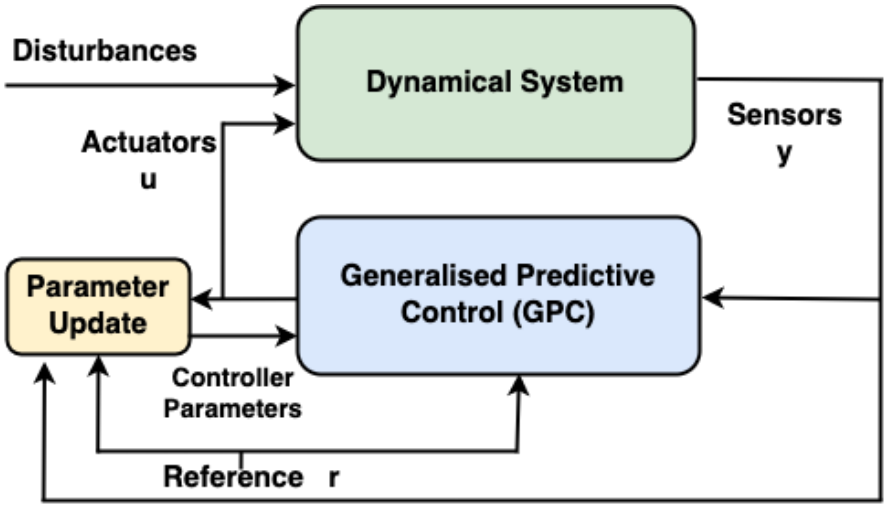


Figure 2.5: Generalised predictive control

Fuzzy Logic [Yen and Langari, 1999] based controllers, which focus on replicating the reasoning of diabetes caregivers, have also been introduced. The MD-Logic System [Atlas et al., 2010] is one such example. Classical control algorithms have been augmented with specialised safety modules, and models to calculate insulin accumulated in the body to adjust insulin infusion to design fully automated systems [Turksoy et al., 2013d; Colmegna et al., 2018]. The performance of fully automated APS research using classical control algorithms based on conducted clinical trials is presented in Table 2.4.

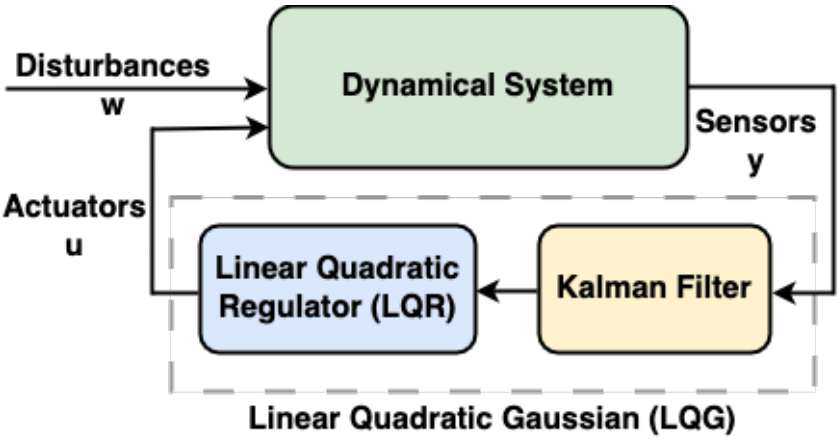


Figure 2.6: Linear quadratic gaussian

The complexity, uncertainty, and disturbances associated with glucose control can be identified as key technical limitations towards the development of fully autonomous APS. This thesis focuses on two approaches towards tackling this challenge. First, RL, an ML-based control approach, is explored for control algorithm

Table 2.4: Performance of fully automated artificial pancreas systems based on clinical trials.

Research Study	Control Algorithm	TIR
Turksoy et al. [2013d]	GPC	46.5%
Turksoy et al. [2013b]	GPC	62%
Turksoy et al. [2014]	GPC	58%
Turksoy et al. [2018]	GPC	76.75%
Colmegna et al. [2018]	LQG	74.7%

Acronyms: GPC - Generalised Predictive Control, LQG - Linear Quadratic Gaussian, TIR - Time In Range.

development. Second, the integration of additional signal sources to design MAPS is explored. The integration of wearable and other invasive sensors is expected to provide additional information related to the glucoregulatory system and daily events affecting glucose control.

2.4 Technical Challenges in Glucose Regulation

There are several technical challenges related to glucose regulation in T1D. These challenges compromise APS performance and usability, and their resolution will lead to safer APS with less cognitive burden on people with T1D. This section discusses the technical challenges of complex dynamics, uni-directional control, delays, uncertainty, disturbances, and variability and their effect on APS.

Complex Dynamics, Uncertainty, and Disturbances. The glucoregulatory system of the human body is a complex dynamical system affected by the disturbances and uncertainty associated with daily events (e.g., meals, exercise, stress, and illness), which challenges glucose regulation [Kovatchev et al., 2009; Cinar, 2017]. Current systems cannot recognise these events and take timely action to maintain normal glycaemic levels. In addition to the activities identified above, natural perturbations, circadian variability, and other biological factors also affect physiological parameters such as insulin sensitivity which affects the performance of APS [Gondhalekar et al., 2013; Stewart et al., 2016].

Sensor and Actuator Delays. Delays are significant limitations in the current hardware, which pose great challenges to APS development. The subcutaneous route used in APS introduces a time delay between the subcutaneous and plasma glucose measurements (roughly 6–12 minutes) and an insulin action delay (roughly 5–15 minutes depending on the type of insulin) [Villena Gonzales et al., 2019; Vliebergh et al., 2021]. These delays are significant during activities such as exercise and meals and affect the optimal control of glucose [Zaharieva et al., 2019]. Moreover, delays in the onset and offset in action when insulin is delivered subcutaneously mean that

such events require timely responses or may even need to be anticipated, resulting in an increased risk of hypoglycaemia and hyperglycaemia [Vliebergh et al., 2021]. In order to address these limitations, researchers are currently focused on developing fast-acting insulin and much more accurate CSGM devices. The focus on sensor failures and noise has also been increased to ensure the development of safe APS [Villena Gonzales et al., 2019].

Uni-directional Control. The insulin-only systems have limited operation since they are single-directional and can only reduce glucose. This limits the protection against hypoglycaemia. Hence, additional control inputs have been introduced in existing treatment [Peters and Haidar, 2018; Ananda Basu, 2020]. Glucagon and amylin are two other hormones closely associated with the glucoregulatory system [Teixeira and Malin, 2008]. Dual hormonal treatment systems [Peters and Haidar, 2018] additionally incorporate glucagon, and tri hormonal systems which incorporate glucagon and amylin have also been explored [Ananda Basu, 2020]. However, these additional hormones are costly and require further clinical assessments to ascertain their safety of administration to the body. Bioelectronic approaches have also been researched *in-silico* to manipulate neural signals to control insulin sensitivity for better glucose regulation [Güemes et al., 2019].

Inter- and Intra-subject Variability. The glucoregulatory system of individuals is subject to variability due to age and other biological factors (food, exercise, psychological stress, illness, season, menses, time of day, pregnancy) [Gondhalekar et al., 2013; Stewart et al., 2016; Cinar, 2017] and it is important to continuously improve the glucose control strategy to adapt to these changes. Hence, there is a large inter- and intra-subject variability in glucose control performance due to the varying glucose dynamics of individuals [Kovatchev et al., 2009]. At present the personalisation of treatment is done by calculating patient-specific characteristics such as TDI, CIR, and ISF [Nathan et al., 1993; Slattery et al., 2018]. These parameters are closely monitored by clinicians and adjusted periodically. Such adjustments need to be made with care by establishing sufficient safety conditions.

2.5 Clinical Challenges in Glucose Regulation

People with T1D and their caregivers face many daily challenges related to existing treatment. Brew-Sam et al. [2021a,b] have conducted studies to identify existing challenges and needs of people with T1D. This section summarises open clinical challenges associated with glucose regulation.

Diabetes Education. The management of T1D is complicated and requires specific knowledge related to glucose sensor, insulin pump management, training and knowledge on the CHO content of different food and meals, and the calculation of insulin

doses. Furthermore, the availability of different commercial insulin pumps and glucose sensors and their unique features increases the effort required, which has motivated studies to understand the user experience of the technologies [Brew-Sam et al., 2021a].

Cognitive Burden. The current HCL systems require users' input regarding meals and exercise; hence similar to previous treatment methods, there remains a significant cognitive burden on the user. Users need to estimate the CHO content of their meal prior to consumption and inform the system in advance regarding any physical activity and the activity type [Shah et al., 2016]. This limits spontaneity and affects the quality of life of people with T1D [Brew-Sam et al., 2021a]. Nearly 50% with T1D consider CHO estimation as the most burdensome aspect of diabetes self-care and often results in missed insulin doses, incorrect meal and correction bolus calculations [Medtronic, 2019]. This highlights the importance of automating the treatment and reducing the cognitive burden. CHO estimation is a difficult task that requires specific training and knowledge on CHO content of different food and meals and frequently leads to errors [Roversi et al., 2020]. In Brazeau et al. [2013], the average CHO counting error by 50 adult T1D subjects was 15.4 ± 7.8g, and the common practice is to consider an error of $\pm$10g per snack or meal acceptable [Roversi et al., 2020]. Computer vision-based systems utilising smartphones are currently being explored towards CHO estimation [Anthimopoulos et al., 2015].

Performance and Safety. The TIR performance of existing systems is around 70%, which can be further improved (Section 2.3.1). The safety of existing APS is challenged due to daily events. Forgetting to bolus could result in severe hyperglycaemia, and miscalculating the doses could result in severe hypoglycaemia, which is a possibility in current HCL systems. The performance and safety considerations of existing treatment also call for attention towards the trust and psychosocial aspects of existing treatment [Brew-Sam et al., 2021a].

2.6 Machine Learning (ML)

The ambition of creating intelligent machines that think dates back to ancient Greece [Bengio et al., 2017], while the Dartmouth Summer Research Project on Artificial Intelligence (1955) is widely considered as the founding event of **Artificial Intelligence (AI)** as a research field [McCarthy et al., 2006]. A member of this project and Turing Award (1969) winning computer scientist, Marvin Minsky, defined AI as;

> *"... the science of making machines do things that would require intelligence if done by men [humans]."*
> *- Marvin Minsky [Minsky, 1968].*

The development of programmable computers in the recent past has made rapid progress in AI, with many practical applications. The game chess was considered

a grand challenge by a generation of AI researchers, ultimately leading to the development of Deep Blue, a chess-playing supercomputer that beat the world chess champion in 1997 [Campbell et al., 2002]. This was a knowledge-based approach for AI, where the system hard-coded human knowledge and used the computational power of computers [Bengio et al., 2017]. **Machine Learning (ML)** was introduced to enable AI systems to acquire their own knowledge by extracting patterns in complex datasets without specific instructions by analysing and drawing inferences [LeCun et al., 2015]. At present, ML is used in many applications and has achieved significant improvements in image recognition, speech recognition, drug discovery, and genomics [LeCun et al., 2015].

ML tasks are mainly classified into three categories (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning. Supervised learning uses annotated datasets to perform regression and classification tasks by learning a function between the input data and the target output [Kavakiotis et al., 2017]. In unsupervised learning, the hidden structure of unlabeled data is discovered through associate rule mining and clustering [Kavakiotis et al., 2017]. Reinforcement learning [Sutton and Barto, 2018] focuses on decision-making by interacting with a dynamic environment and will be covered in detail in Chapter 3.

Deep Learning (DL) is a subfield in ML based on **Artificial Neural Networks (ANN)** and is defined as a type of ML algorithm where multiple layers of processing are used to learn representations with multiple levels of abstractions from large datasets, which is a limitation in conventional ML techniques [LeCun et al., 2015]. Convolutional Neural Networks (CNN) [LeCun et al., 1989], Temporal Convolution Networks (TCN) [Lea et al., 2016], Recurrent Neural Networks (RNN) [Hihi and Bengio, 1995], and Long-Short Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997] are some of the most popular ANN. DL has been enabled thanks to the advancements in computing (e.g., Graphical Processing Units (GPU), Tensor Processing Units (TPU)).

2.7 Machine Learning in Type 1 Diabetes

The ability of ML to utilise multi-modal data streams, learn complex physiological relationships, personalisation capabilities, and handle uncertainties has resulted in a significant emphasis on the application of ML towards T1D in recent years [Contreras and Vehi, 2018]. These efforts mainly relate to glucose control [Tejedor et al., 2020], biomarker identification [Kavakiotis et al., 2017], blood glucose prediction [Hameed and Kleinberg, 2020], blood glucose dynamics modelling [Woldaregay et al., 2019], detection of adverse glycaemic events [Contreras and Vehi, 2018], and personalised decision support systems [Woldaregay et al., 2019; Kavakiotis et al., 2017]. Several of these research directions are specifically important towards developing APS. In this section, previous research on ML in T1D is summarised based on their applicability to APS (Figure (2.7)). This work can be mainly categorised as (1) the development of auxiliary modules for APS, (2) glucoregulatory system modelling, and (2) control

algorithm development.

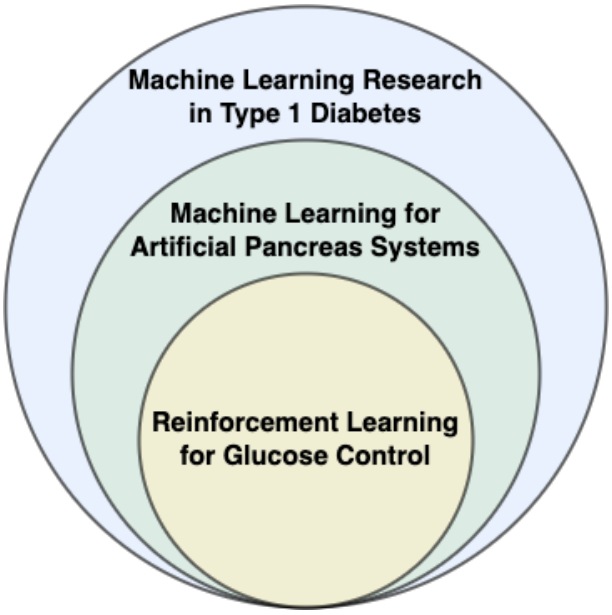


Figure 2.7: Applications of machine learning for type 1 diabetes.

The operation of APS is challenging due to the complexities of glucose regulation and the uncertainty associated with the daily events of the user. Hence, the system requires auxiliary components in addition to the primary control strategy. The designed auxiliary modules are expected to provide the primary control algorithm with feedback to improve the insulin delivery strategy. Such modules have been explored in previous studies aiming to develop hypoglycaemia early alarm systems [Turksoy et al., 2013c], hypoglycaemia prediction [Turksoy et al., 2016], meal detection [Turksoy et al., 2015; Ramkissoon et al., 2018; Hajizadeh et al., 2019], and CHO estimation [Turksoy et al., 2016]. ML techniques can be used to complement some existing systems while improving others through the use of their improved prediction and forecasting capabilities under uncertainty [Bengio et al., 2017]. ML techniques have been extensively used in improving these modules. Vision-based systems and CNNs have been used for CHO estimation systems [Vasiloglou et al., 2018]. RNN, Support Vector Machines (SVM), and Auto-Regression with Exogenous input (ARX) models have also been used for hypoglycaemia prediction algorithms [Tyler and Jacobs, 2020]. The potential for novel auxiliary modules is immense as additional wearable sensors can be used to develop such systems [Hettiarachchi et al., 2022a]. Both invasive and non-invasive sensors can be applied to design modules for exercise detection [Breton et al., 2014; Castle et al., 2018] and diabetes ketoacidosis detection [Lee et al., 2019], which are expected to be valuable in glucose control. Focusing on such auxiliary modules will help design comprehensive APS since they are expected to contribute to the safety, personalisation and explainability characteristics.

Control engineering and ML fields share overlapping ideas and methods. **System**

Identification (SI) in control engineering focus on learning a model of the underlying dynamical system [Yassin et al., 2013]. Methods in SI can be categorised as (1) white-box approaches which derive models from first principles; (2) black-box approaches based on data acquired from the system; and (3) grey-box approaches, which utilise prior knowledge about the system and data [Yassin et al., 2013]. The current glucoregulatory system models used in T1D are predominantly based on grey-box approaches, which use compartmental models and are personalised using data [Wilinska et al., 2010; Visentin et al., 2018]. Previous research has also focused on black-box approaches where subspace identification [Hajizadeh et al., 2017] and ARX [Turksoy et al., 2013b] methods are used for SI. Utilising the capabilities of ML in learning complex relationships, techniques such as Random Forest [Liaw et al., 2002], SVM [Xuegong, 2000], ElasticNet [De Mol et al., 2009], and Gradient Boosting Trees [Ke et al., 2017] and DL models such as LSTM [Hochreiter and Schmidhuber, 1997] networks, TCN [Lea et al., 2016] have been applied towards data-driven glucoregulatory system modelling [Xie and Wang, 2020, 2018]. Xie and Wang [2020] presents a comparison study between these ML approaches and classical time series models. It is important to note that various multi-inputs from wearable devices, in addition to glucose and insulin, have also been used towards system modelling in order to improve the performance [Rashid et al., 2019].

Real-world control problems are challenging due to their high-dimensional, non-linear complex dynamics, which results in difficulty in system modelling and developing useful control algorithms [Duriez et al., 2017; Brunton and Kutz, 2022]. Hence, data-driven approaches to control are being explored, and the power of ML algorithms has enhanced the focus on **Machine Learning Control (MLC)** [Duriez et al., 2017]. Various approaches towards MLC have been investigated. As identified above, ML is used for SI, where classical control algorithms leverage the model for designing the control algorithm [Duriez et al., 2017]. ML has also been used to tune parameters of classical control algorithms such as PID control [Duriez et al., 2017]. Techniques such as genetic programming [Koza et al., 1999] which is an evolutionary algorithm, and adaptive neural networks [Palnitkar and Cannady, 2004] have been used to design the structure and tune parameters of control algorithms [Duriez et al., 2017]. These approaches use an **optimal control** framework where the goal is to control the dynamical system by optimising an objective function, using data in an off-line learning loop (Figure 2.8) [Duriez et al., 2017].

RL techniques are also used for control, where a control strategy is improved over time to achieve a defined objective through experience. A majority of MLC approaches in APS development are related to RL. In contrast to other ML methods, RL is more suited for decision-making tasks in a dynamic system environment. Also, RL algorithms have been shown to perform well under unknown variable delays (through delayed reward mechanisms); in learning complex non-linear dynamics; handling uncertainties and disturbances; and in personalisation [Bothe et al., 2013]. Hence, RL-based algorithms are currently being explored in glucose control in T1D [Daskalaki et al., 2013a]. An introduction to RL and a review of previous research on RL-based APS are presented in Chapter 3.

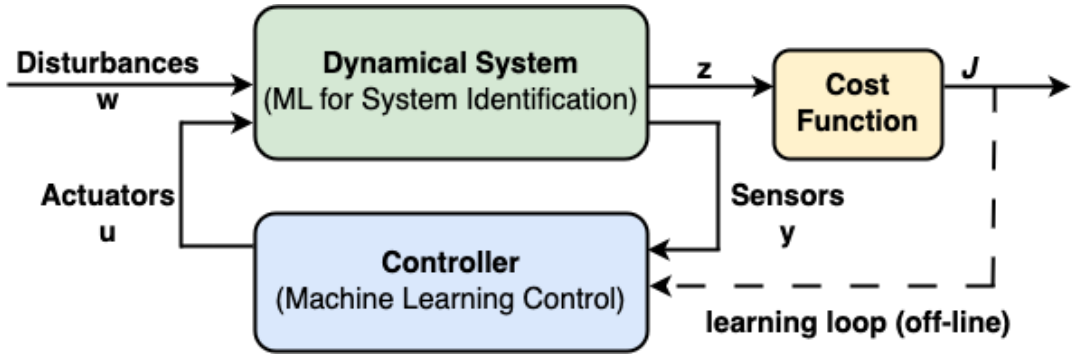


Figure 2.8: Applying machine learning in the standard framework for feedback control. The system is affected by disturbances $\boldsymbol{w}(t)$ and controlled by noisy actuators $\boldsymbol{u}(t)$ based on noisy system measurements $\boldsymbol{y}(t)$. The control objective is to minimise a cost function J based on vector $\boldsymbol{z}(t)$, which contains required information related to the system. The off-line learning loop provides data to train the controller.

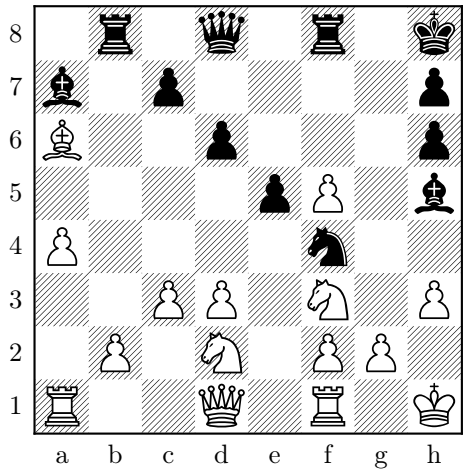
2.8 Summary

This chapter introduced the problem of glucose regulation in T1D and presented the evolution of T1D treatment methods. HCL systems were identified as an emerging treatment strategy based on CSGM and CSII, where PID and MPC algorithms have been used for control algorithm development. The potential for fully autonomous APS was identified. However, there are many challenges associated with the problem of glucose regulation. The cognitive burden on people with T1D and their carers was identified as a major limitation of existing treatment, while problem-specific challenges such as complex dynamics, uncertainty, disturbances, uni-directional control, delays, and high inter- and intra-subject variability are also present. The next chapter presents the latest research on using RL algorithms for glucose control and their suitability to address identified challenges.

Chapter 3

Background and Related Work II: Reinforcement Learning

Chess is the most widely-studied domain in the history of artificial intelligence. In 2017, AlphaZero a RL-based chess program, achieved a superhuman level of play without being taught any chess strategies by its human creators. AlphaZero is capable of discovering novel strategies, as shown in the following game between Stockfish 8 vs AplhaZero (0-1), London 2018(12), where it sacrifices material and king side structure to mobilise its pieces [Sadler and Regan, 2019].



Stockfish 8 vs AlphaZero
London 2018(12) (0-1)

This chapter provides essential background material required to develop RL-based APS. First, RL is introduced along with its concepts and applications. Next, challenges associated with applying RL to real-world problems are presented. Finally, a literature review on the use of RL for continuous control problems and APS is presented, and limitations of existing RL-based APS are highlighted to identify areas of improvement.

"Chess has been shaken to its roots by AlphaZero, but this is only a tiny example of what is to come. Hidebound disciplines like education and medicine will also be shaken, if slowly, by the improved results promised by AI analysis, if we allow them to."
- Garry Kasparov
(World Chess Champion, 1985-2000)
[Sadler and Regan, 2019, p.16].

3.1 Reinforcement Learning (RL)

The field of **Reinforcement Learning** is concerned with learning how to act in an environment to achieve a specified objective. The origins of RL are mainly from psychology and optimal control, which have been pursued independently [Sutton and Barto, 2018]. The "Law of Effect" in psychology introduced learning by trial and error, which described the effect of reinforcing events based on the tendency to select actions [Thorndike, 1911]. The underlying theory and algorithms in optimal control were also essential in developing RL. A **stochastic optimal control** problem in optimal control deals with uncertainties present in the system and is formulated as a Markov Decision Process (MDP). Dynamic programming is used to solve these problems. However, it suffers from the curse of dimensionality, where computational requirements grow exponentially with the number of state variables [Sutton and Barto, 2018]. RL focus on problems related to optimal control and, specifically, on stochastic optimal control problems. Integrating learning by trial and error and optimal control concepts around temporal difference methods [Samuel, 1959] in the late 1980s led to modern RL [Sutton and Barto, 2018]. Temporal-difference methods focus on deriving predictions for a quantity based on temporally successive estimates of the given quantity through bootstrapping. It was successfully used in a learning algorithm to develop a checkers-playing program [Samuel, 1959]. At present, RL can be identified as an intertwining of many fields such as ML, control, psychology, neuroscience, mathematics, economics, and engineering [Sutton and Barto, 2018].

Concepts and Terminology. RL is a computational approach where a decision-maker, called an **agent**, interacts with an **environment** to learn a strategy (**policy**) to achieve a specific goal. The interaction between the agent and environment is formulated using the formal MDP framework [Sutton and Barto, 2018]. A policy maps each **state** (s_t) of the environment to an **action** (a_t). At a given state, the agent takes an action, which results in a state transition and a feedback called a **reward** (r_t), which evaluates the transition related to the goal pursued. The agent uses its **experiences** ($s_1, a_1, r_1, s_2, a_2, r_2 \cdots$ transitions) to learn a policy to maximise the expected cumulative reward **(return)** from any given initial state. The expected return for a given state is also called the **value**. Unlike other approaches, RL emphasises learning through direct interaction with the environment without requiring supervision or complete models of the environment [Sutton and Barto, 2018]. A formal definition is presented in Chapter 4.

Model-based and Model-free RL. Model-based RL algorithms rely on learning a model of the environment in order to optimise their control policy. They are more sample-efficient than model-free methods and benefit from generalising well to new tasks and environments [Bansal et al., 2017]. On the other hand, they heavily depend on the accuracy of the learnt model [Xiao et al., 2019]. Model-free RL algorithms have outperformed model-based algorithms in domains where model learning is challenging. However, they require millions of trials for learning [Bansal et al., 2017;

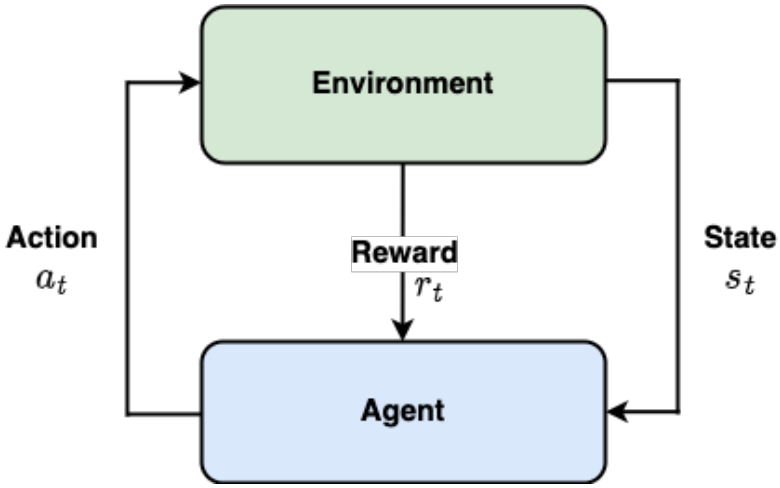


Figure 3.1: Reinforcement learning framework. At each time step, an agent interacts with an environment by taking an action (a_t) based on the environment state (s_t) and receives a feedback called a reward (r_t).

Schrittwieser et al., 2020]. Several prior works have focused on combining model-free and model-based paradigms to design RL systems [Sutton, 1991; Gu et al., 2016; Bansal et al., 2017; Schrittwieser et al., 2020].

On-policy and Off-policy RL. RL algorithms can also be classified as on-policy and off-policy algorithms. In on-policy learning, the policy which is being learned is also used for action selection. In contrast, in off-policy learning, a different policy is used for action selection (called a behaviour policy). Off-policy algorithms are more sample efficient due to the ability to re-use collected experiences, while on-policy algorithms are more stable and easy to use [Gu et al., 2017].

Exploration and Exploitation in RL. In order to learn a suitable policy, the RL agent should sufficiently explore different strategies for better action selection and also exploit already learned strategies to increase its reward. Hence, the trade-off between exploration and exploitation is vital in RL.

3.2 Reinforcement Learning Applications

Board games are the most popular application field in Modern RL. It started gaining attention through the development of the TD-Gammon program in 1992, which achieved the Master Level in the game of Backgammon [Tesauro, 1994]. Next, RL was applied to the game of Chess, where the program DeepBlue defeated the world chess champion in 1997 [Campbell et al., 2002]. The introduction of DL accelerated RL and enabled scale to problems previously thought intractable [Arulkumaran et al., 2017]. This led to the focus on more complicated and challenging games such as Go and computer games such as Atari, which contained games like Pong, Breakout, and

Space Invaders. Deep Reinforcement Learning (DRL) algorithms were designed to play Atari games using display pixel input to achieve human expert performance [Mnih et al., 2013]. In 2016, the algorithm AlphaGo [Silver et al., 2016] beat the world Go champion. Since then, advanced algorithms such as AlphaZero (2017) [Silver et al., 2017], which achieves a superhuman level of playing in games (Chess, Shogi, Go) solely by self-play without handcrafted evaluation functions, and MuZero (2020) [Schrittwieser et al., 2020] which achieves superhuman performance without the use of any knowledge of the game dynamics (i.e., rules of the game) have been introduced. The latest research focus is on imperfect information games such as Stratego (DeepNash (2022) [Perolat et al., 2022]) and Multi-Agent Reinforcement Learning (MARL), where algorithms are designed to play multi-player games such as StarCraft (AlphaStar (2019) [Vinyals et al., 2019]) and No-press Diplomacy (Diplodocus (2022) [Bakhtin et al., 2022]).

The application of RL is explored in continuous control problems [Schulman et al., 2017; Haarnoja et al., 2018], intelligent transportation systems [Haydari and Yılmaz, 2020], commercial cooling systems [Luo et al., 2022], and healthcare applications [Liu et al., 2020]. Continuous control has gained much attention due to applications such as 3D humanoid motion problems, self-driving, and dexterous manipulation tasks [Lillicrap et al., 2015; Haarnoja et al., 2018; Kendall et al., 2019; Akkaya et al., 2019]. RL applications in healthcare include glucose control in T1D [Daskalaki et al., 2013a], propofol dosing in general anaesthesia [Vajapey, 2019], and multi-cytokine therapy for sepsis [Petersen et al., 2018]. Reinforcement Learning through Human Feedback (RLHF) approaches was also recently used to fine-tune large language models such as instructGPT and chatGPT [Ouyang et al., 2022]. The following section presents the challenges associated with these real-world RL applications.

3.3 Challenges in Real-world Reinforcement Learning

Significant technical challenges are associated with using RL algorithms in real-life applications [Dulac-Arnold et al., 2019, 2020]. Hence, it is crucial to identify these limitations before applying RL to the safety-critical application of glucose control. This section analyses these challenges to help design improved practical systems.

Complex Environments. Real-world environments are complicated compared to game environments, as they could be only partially observable, stochastic, with noisy sensor/actuator systems with delays, and non-stationary. These problems are formulated as Partially Observable MDP (POMDP), where either the history is incorporated in the agent [Mnih et al., 2015] or RNNs used within the agent [Hausknecht and Stone, 2015]. The modelling of the system as non-stationary with a time-varying reward function was explored in Nagabandi et al. [2018a]. System delays have been addressed by incorporating recent history in agents [Hester et al., 2018], using memory-based agents [Hung et al., 2019], and using a backward-view of a task to generate

the return-equivalent MDP [Arjona-Medina et al., 2019].

Reward Function Design. RL agents are optimised to achieve a specified goal which is formulated based on a reward function. The reward function is trivial in simple tasks. However, in real-world problems, the formulation of reward functions is much more complicated. In some tasks, a clear reward might not be present, requiring domain experts and RL practitioners to work closely to formulate reward functions. Tasks may also be more complex due to the existence of multiple objectives. Previous work has explored Inverse Reinforcement Learning (IRL) approaches to recover an underlying reward function from demonstrations [Ng et al., 2000; Abbeel and Ng, 2004]. However, in some problems, the expert demonstration may not be perfect (e.g., existing clinical treatment strategies for glucose regulation).

High-dimensional Continuous State and Action Spaces. Real-world problems are often associated with large, continuous state and action spaces compared to tasks such as games which contain small discrete action spaces. This results in challenges related to the efficient exploration [Lazaric et al., 2007] and the existence of redundant or irrelevant actions [Zahavy et al., 2018], which makes it hard to learn a suitable strategy. In previous work, Lazaric et al. [2007] used a novel actor-critic algorithm based on a Sequential Monte Carlo approach to improve exploration, while Zahavy et al. [2018] proposed an approach that combined the RL algorithm with an action elimination network to eliminate sub-optimal actions.

Safety Requirements. Safety is paramount in real-world systems, especially in healthcare applications concerning vital core functions such as glucose regulation, where patients' lives may be at stake. RL algorithms are required to sufficiently explore potential strategies to learn a suitable strategy for the underlying task. However, this process could lead to catastrophic failures. Previous research on safe exploration has focused on Constraint MDPs [Dalal et al., 2018; Achiam et al., 2017], Probabilistic Goal MDPs [Mankowitz et al., 2016], Lyapunov functions [Turchetta et al., 2016; Chow et al., 2018; Wachi et al., 2018], additional safety layers for systems [Dalal et al., 2018; Pham et al., 2018], and the use of human feedback for avoiding failures [Saunders et al., 2017]. In some applications, safety violations during training may be acceptable (e.g., using simulators for training RL algorithms before deployment). For such applications, methods that can enforce safety conditions for trained policies have been proposed [Achiam et al., 2017; Tessler et al., 2018; Bohez et al., 2019].

Explainability. The learned policies of agents are complicated and hard to explain. However, when applying learned policies to the real world, the users and operators of the underlying systems require assurances of the operation and insights into the policy. Hence, explainability is of utmost importance. Previous work has explored model-based methods, which use rollouts for planning to provide insights into the policy [Chua et al., 2018; Hafner et al., 2019].

On-line / Off-line Learning and Transfer Learning Constraints. Typically, RL agents require millions of interactions towards learning. Through these interactions, RL agents obtain the experiences required for learning (Section 3.1). Hence, learning on-line (i.e., *in-vivo*, on-patient) may be infeasible in some real-world applications due to the system's slow-moving nature and fragility (For example, in APS, insulin infusion actions are typically made every 5-minutes, which would require 9.5 human years to complete 1 million interactions, while safety constraints would limit the quality of the experience). Furthermore, it could be expensive to collect experiences, and the quality of experiences might be low. Previous work has explored expert demonstrations [Mnih et al., 2015] and improving sample-efficiency of algorithms to overcome this challenge [Finn et al., 2017]. However, when expert demonstrations are unavailable, simulators are required to train and develop data-efficient agents. Furthermore, the simulators enhance the exploration capabilities of RL agents. However, the simulators used in the target problems may not be perfect, which presents a challenge of safely transferring the learnt policy and continually learning on the real system for fine-tuning. Finn et al. [2017] explored using few-shot learning to efficiently adapt to new in-distribution tasks.

As an alternative to the online learning approach, off-line off-policy learning approaches, which use existing system logs, can be explored. This is challenging when the difference between the behaviour policy (in which the data was collected) and the target policy (learned by the agent) is large, which results in different state distributions. Importance sampling [Precup et al., 2000], the direct method which uses a learned transition model for evaluation, and doubly robust estimators [Dudík et al., 2011; Jiang and Li, 2016] have been introduced to address this challenge. These practical constraints must be carefully evaluated to design a feasible RL system.

3.4 Reinforcement Learning for Continuous Control

Glucose regulation in T1D is a continuous control problem. Hence, this section presents RL techniques developed for continuous control tasks. The use of RL for continuous control has gained much attention due to applications such as locomotion, self-driving, and dexterous manipulation tasks [Lillicrap et al., 2015; Haarnoja et al., 2018; Akkaya et al., 2019; Kendall et al., 2019]. Both on-policy (Advantage Actor Critic (A2C) [Mnih et al., 2016], PPO [Schulman et al., 2017]) and off-policy (Soft Actor Critic (SAC) [Haarnoja et al., 2018]) algorithms have been used in model-based [Nagabandi et al., 2018b] and model-free [Mnih et al., 2013] schemes. PPO [Schulman et al., 2017] is a widely used model-free on-policy algorithm that has shown promise in many RL tasks [Andrychowicz et al., 2021]. It is an actor-critic method consisting of policy and value functions. The optimisation objective of the policy function is to learn a stochastic policy, while the objective of the value function is to learn the value of the system state. In standard on-policy policy gradient methods, past data cannot be used to improve the current policy. Hence, they are sample-inefficient. The

PPO algorithm improves the sample efficiency by using a clipped objective function, which enables multiple updates for a given data sample. This objective also avoids excessive changes to the policy while training. PPO is implemented using either a shared neural network for the policy and value functions or separate neural networks [Cobbe et al., 2021]. The former facilitates feature sharing between the two functions, while the latter avoids interference between the two optimisation objectives.

Auxiliary Learning. The Phasic Policy Gradient (PPG) algorithm [Cobbe et al., 2021] is a variant of PPO, which uses separate neural networks and introduces an additional auxiliary learning task of value function estimation. The auxiliary learning facilitates the distillation of features between the two networks and further improves the sample efficiency. Similar ideas based on auxiliary learning tasks have been explored in deep Q-learning approaches where the sample efficiency and policy adaptation during deployment were improved [Hansen et al., 2020; Schwarzer et al., 2020]. Hence, auxiliary learning tasks are increasingly being used in deep RL algorithms. Model learning is a popular auxiliary learning task useful in updating policy and value functions through planning and action selection [Hessel et al., 2021]. Previous work in RL carries out auxiliary model-learning by predicting latent state representations [Lee et al., 2020a; Schwarzer et al., 2020], predicting observations/system states [Bansal et al., 2017; Schrittwieser et al., 2020], and through estimating future rewards, value and policy functions [Hessel et al., 2021].

Planning. A learnt model of the environment can be used for simulations and planning. *Planning* is a process that uses a learned model to improve the policy, where the RL algorithm interacts with the modelled environment. The two distinct approaches to planning are state-space planning and plan-space planning [Sutton and Barto, 2018]. In state-space planning, the RL algorithm uses the model to simulate experiences, which are then used to update the value function and, ultimately, the policy function. Simulating experiences is valuable when real experiences are costly and limited. Dyna-Q [Sutton, 1991] is such an algorithm where the experiences of the RL algorithm are used for both model-learning and planning in an online manner. In plan-space planning, the planning is conducted as a search over the space of plans (e.g., partial order planning). Previous work has introduced such methods based on Monte Carlo Tree Search (MCTS), where an expert policy is used for planning [Anthony et al., 2017]. The planning horizon reflects the objective of the control task. Most real-life control problems often have both short-term and long-term objectives. Designing reward functions to capture both objectives is challenging or even infeasible [Khorasgani et al., 2019]. Khorasgani et al. [2019] have proposed a model-based RL approach based on the Q-learning algorithm to focus on the two objectives by modelling short/long-term prediction models. In the discounted RL setting, a discount factor (γ) is used to optimise an exponentially decreasing function of the future return. The magnitude of γ establishes an effective horizon for optimising the agent [Sutton and Barto, 2018]. Several prior studies have also focused on learning value functions over multiple time horizons [Sutton, 1995; Fedus et al., 2019; Romoff

et al., 2019] to address short/long-term objectives.

3.5 Reinforcement Learning based Artificial Pancreas Systems

In the problem of glucose control, the goal can be formulated based on the clinical objective to improve TIR. The agent is called to make decisions regarding the continuous insulin infusion rate. The environment refers to a person with T1D (either real or *in-silico*), their management conditions (e.g., using a CSGM, CSII pump with resulting delays), and their external disturbances (e.g., meals, exercise) (Figure 3.2). The agent's observations relate to continuous glucose information and other parameters that are potentially monitored (e.g. ketones, heart rate), and the actions refer to the infusion of insulin, as well as other potential interventions that may be considered (e.g. glucagon infusion). After an action is performed, the agent receives feedback, reflecting the control strategy's quality based on the formulated goal.

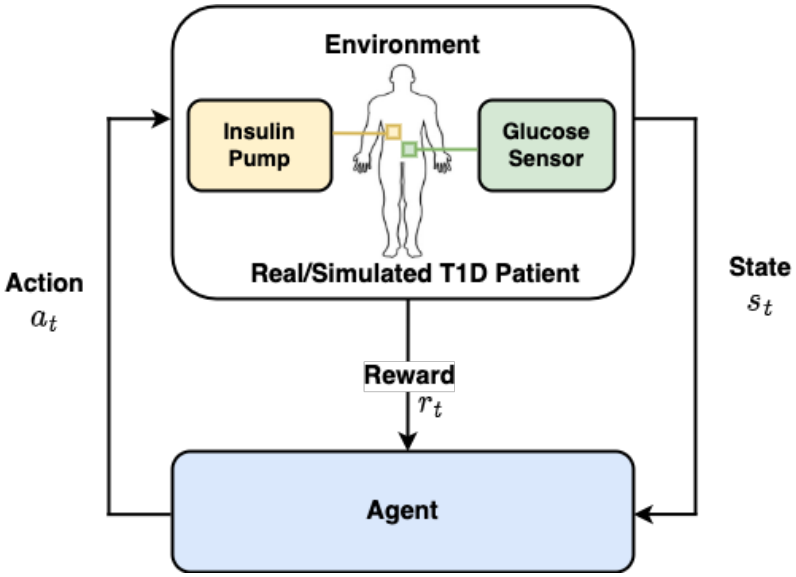


Figure 3.2: Basic structure of an RL setting in glucose control.

3.5.1 Literature Review

The pioneering study on using RL for glucose control in T1D dates back to 2009, where Yasini et al. [2009] explored Q-Learning for controlling hyperglycaemia [Bothe et al., 2013]. In 2012, Daskalaki et al. [2013a,b] explored the use of Actor-Critic methods. However, due to the limitations of the available simulators, the task was restricted to daily updates to the control strategy [Daskalaki et al., 2013b; Sun et al., 2018, 2019]. The development of the UVA/PADOVA (2014) simulator [Dalla Man et al., 2014] and the Simglucose (2018) simulator [Xie, 2018] has resulted in recent studies where insulin updates have been possible in near real-time. In the

UVA/PADOVA simulator, the control algorithm must be designed within a Matlab Simulink block, which limits the application of RL. Hence, Simglucose [Xie, 2018] was designed as an open-source python implementation of the FDA-approved UVA/PADOVA (2008) simulator. The Simglucose simulator can leverage the existing PyTorch [Paszke et al., 2019] machine learning framework to develop control algorithms for APS and allow reproducibility of proposed algorithms. This also provided the capability to easily and efficiently experiment with advanced RL algorithms (e.g., algorithms based on deep neural network architectures and using parallel agents during training). The simulators comprise adult, adolescent, and children T1D cohorts with ten subjects each, representing patient variability found in a real T1D population and allowing for simulating different meal scenarios. Section 4.2.1.1 provides an in-depth overview of the simulators.

The studies which propose RL-based APS can be classified as (1) hybrid systems which only control basal insulin and (2) fully automatic APS. A majority of the previous work focused on hybrid approaches where Q-learning [Myhre et al., 2018; Ngo et al., 2018a,b] and Deep Q-learning (DQN) [Zhu et al., 2019, 2020; Fox and Wiens, 2019] was used. Q-learning is one of the oldest algorithms in RL, and integrating DL has resulted in DQN. As Q-learning-based approaches require discrete actions, these studies used handcrafted discrete actions of the insulin infusion rate. The SAC algorithm [Haarnoja et al., 2018], which focused on a continuous action space, was introduced in 2018 and Lim et al. [2021] proposed a system based on SAC to estimate insulin, guided by a PID controller. The system also used a random forest regressor and a dual attention network for glucose prediction. All of the above systems require meal announcement, CHO estimation or both. Hence they cannot eliminate the cognitive burden on the users associated with the tasks.

Recent studies have focused on fully automatic RL-based systems for glucose regulation. Fox et al. [2020] proposed a SAC-based system that used historical glucose and insulin values as input for glucose control. However, the system only focused on daily meals with low CHO contents. Lee et al. [2020b] proposed a bio-inspired RL approach using the PPO algorithm [Schulman et al., 2017] proposed in 2017, where reward functions and discount factors of the algorithm reflected subject-specific pharmacological characteristics and temporal homeostatic objectives (e.g., discounted rates derived from individual pharmacokinetics and pharmacodynamics (PK/PD)). The system was only evaluated based on the *in-silico* adult T1D patient population of the simulator. Both Fox et al. [2020] & Lee et al. [2020b] aims to eliminate CHO counting. Table 3.1 summarises these fully autonomous RL-based systems and includes the recent hybrid system proposed by Lim et al. [2021] for comparison. A comprehensive review of RL in glucose control is presented in Tejedor et al. [2020].

Table 3.1: Performance of reinforcement learning based control systems.

Study	System	Performance†		Simulator
		Adults	Adolescents	
Fox and Wiens [2019][a] SAC	FA	Full cohorts not considered		Simglucose 2018 (UVA/PADOVA 2008)
Fox et al. [2020][b] SAC (RL-MA) SAC (RL-Scratch)	Hybrid* FA	77.12 72.68		Simglucose 2018 (UVA/PADOVA 2008)
Lee et al. [2020b] PPO	FA	89.30±4.19	-	Custom platform based on UVA/PADOVA 2014
Lim et al. [2021] SAC	Hybrid*	65.93±17.29	62.20±19.99	Custom platform based on UVA/PADOVA 2014

†**Metrics**: The median TIR is presented for Fox et al. [2020], while TIR presented as mean±std for Lee et al. [2020b] and Lim et al. [2021].
†**Meal Protocol**: Fox and Wiens [2019] and Fox et al. [2020] used low CHO meals (values not provided) while Lee et al. [2020b] and Lim et al. [2021] used meals with average daily CHO of 180g.
[a] Experiments were conducted only on three handpicked subjects, [b]Results presented as the median TIR across all cohorts and is the only RL-based study which includes the Child cohort. *Require at least meal announcement.
Acronyms: CHO - Carbohydrates, FA - Fully Autonomous, MA - Meal Announcement, PPO - Proximal Policy Optimisation, RL - Reinforcement Learning, SAC - Soft Actor Critic, TIR - Time In Range.

3.5.2 Limitations and Areas of Improvement

Only a handful of RL-based studies have focused on designing fully autonomous systems. RL-based APS designs need to focus on realistic meal scenarios with large CHO contents to stress-test the system, which is a limitation in existing research [Fox and Wiens, 2019; Fox et al., 2020]. In addition, there is a large variability among people with T1D; specifically, the adolescent patient cohort is more challenging to control [Kovatchev et al., 2009]. Hence, proposed systems must be validated on all subjects across multiple *in-silico* cohorts to evaluate their performance, which is a limitation of some previous work [Lee et al., 2020b]. It is also essential to focus on generalised formulations, as methods that over-personalise based on *in-silico* subjects may not be translatable to real life [Lee et al., 2020b]. Therefore, future research efforts in RL-based APS are vital to developing generalised methods for handling patient variability and realistic meal scenarios.

Applying RL to APS must contend with real-world challenges identified in Section 3.3. The glucoregulatory system is a highly complex biological system with uncertainties, disturbances, and unknown delays in glucose sensing and insulin action. It is a safety-critical application that requires expert clinical knowledge for

reward function formulation and faces challenges related to explainability and real-world learning constraints. The problem also is associated with large state and action spaces. The aspects of safety, explainability, and translation to real life in RL-based APS need to be further explored. These characteristics must be built into the problem formulation and RL algorithm design. This can be achieved by focusing on catastrophic failures, stability, architecture, and efficiency of RL algorithm design.

RL is an emerging and rapidly evolving area of research. Developing fully automated RL-based APS requires coordinating multidisciplinary teams involving people with T1D, health experience experts, clinicians, engineers, and computer science researchers. Hence, bridging the communication barriers between these stakeholders is vital. The black-box nature of ML methods restricts the explainability of the control strategies [Tjoa and Guan, 2020]. Existing static treatment strategies are easy to understand due to the distinct insulin infusion patterns related to meals and background insulin actions. However, the RL-based glucose control strategies are more complex, with varying insulin levels. Current APS based on PID and MPC control strategies presents tunable parameters which clinicians understand (e.g., glucose target, insulin factor) [Steil et al., 2011; Bequette, 2013]. Hence, RL-based APS could focus on designing tunable parameters with clinical feedback while focusing on underlying biological concepts of the glucoregulatory system [Kovatchev et al., 2009]. These parameters would help with explainability and are expected to help clinicians experiment with RL-based APS and understand their operation. The feedback would also be valuable to design RL-based APS to cater to the application's specific needs and thus could result in better performance. Hence, explainability can be viewed as both the ability of the engineer to understand the clinical requirements as well as the ability of the clinicians to understand the control strategy learnt.

3.6 Summary

This chapter introduced RL concepts, terminology, and applications. RL challenges towards real-world applications were identified, and previous work on RL-based APS for glucose control was explored to identify existing limitations and areas of improvement. The following chapters of this thesis focus on presenting an RL-based problem formulation, RL algorithm design, and approaches to address identified challenges.

Chapter 4

Problem Formulation and State-Of-The-Art RL for Glucose Control

In previous chapters, clinical and technical challenges of APS were discussed. The cognitive burden on people with T1D and their carers was identified as an important area for improvement.

"Pretty much constantly managing it. I weigh her food. I have a carb scale that tells the carbs. So for every meal I'm weighing things, telling her how much insulin to give and on her CGM [...] So I'm constantly looking at my phone to see what her blood sugar is."
[Hughes et al., 2020, p.913]

Existing APS are hybrid hybrid closed-loop systems which requires meal announcement and CHO estimation. Closed-loop systems have the potential to improve glucose regulation and are also requested by people with T1D to minimise the user burden as identified by the study conducted by Brew-Sam et al. [2021b].

"What is needed is an all in one device (CGM, insulin pump and control system) that doesn't require tubes and can be controlled via an app with an algorithm that constantly regulates blood sugars that can operate as a closed loop system."
[Brew-Sam et al., 2021b, p.11].

This chapter presents a closed-loop RL-based APS design, targeting the cognitive burden associated with existing treatment related to CHO estimation and meal announcement. First, a formulation for the glucose control problem is proposed by focusing on the environment, state/action spaces, and reward function design. Next, the state-of-the-art RL algorithms are explored and implemented to compare against standard clinical treatment based on basal-bolus treatment strategies. Finally, the experimental design and evaluation metrics are presented, and the performance of the designed fully-automated RL-based APS are evaluated.

4.1 Introduction

Existing T1D treatment adds a significant cognitive burden to people with T1D and their carers. ML techniques have been explored in different aspects of treatment to alleviate this burden, as identified in Chapters 2 and 3. Eliminating meal CHO estimation and meal announcement and automating correction insulin doses provides an opportunity to dramatically reduce a key contributor to the cognitive burden on T1D users. This chapter explores the use of RL towards fully automating T1D treatment. As identified in Chapter 3 the majority of proposed RL-based APS to date require manual intervention. Only two fully automated systems have been proposed [Lee et al., 2020b; Fox et al., 2020].

Formulating the glucose control problem considering the application-specific and technical requirements is essential to developing RL-based control algorithms. The problem formulation mainly focuses on the state space, action space, and reward function designs. The RL-based approaches proposed in previous work have formulated the glucose problem in a variety of different ways. The state space and underlying MDPs of the RL problem are also formulated using different approaches [Lee et al., 2020b; Fox et al., 2020; Zhu et al., 2019, 2020]. Some studies focus on handcrafted discrete action space [Zhu et al., 2019, 2020; Fox and Wiens, 2019], while others focus on continuous action spaces for insulin [Fox et al., 2020]. Also, a variety of different reward function formulations have been proposed based on clinical risk indices [Fox and Wiens, 2019; Fox et al., 2020], sparse reward formulations [Zhu et al., 2019, 2020], and experimental approaches [Lee et al., 2020b]. Section 4.2.1 of this chapter presents a novel formulation of the glucose problem, which aims to address the challenges of fully automating T1D treatment.

In recent years, many RL algorithms have been proposed and tested on various problem domains. In Section 4.2.2 of this chapter, the state-of-the-art algorithms A2C (2016) [Mnih et al., 2016], PPO (2017) [Schulman et al., 2017], and SAC (2018) [Haarnoja et al., 2018] are explored to evaluate their suitability to design RL-based APS. These algorithms are selected due to their applicability to continuous control problems, which focus on continuous action spaces. Following established best practices in the field, clinical benchmarks, experimental protocols, and evaluation metrics are identified to compare the proposed methods. Finally, the characteristics of the candidate algorithms are analysed based on their clinical performance and learning behaviour.

4.2 Methodology: Eliminating Meal Announcement and Estimation

The methodology is divided into problem formulation and RL algorithm design. In the problem formulation section, the glucoregulatory system is identified as a dynamical system, and the environment is defined based on state/action space and reward function designs. The RL algorithm section identifies suitable algorithms for

the task and presents their derivations, design, and implementation details.

4.2.1 Problem Formulation

This section first introduces the glucoregulatory system, available simulator models, and the simulator selected in this thesis. Next, the environment of the glucose control problem is formulated, and its characteristics are highlighted. Finally, designs for the state/action space and reward function are presented.

4.2.1.1 The Glucoregulatory System

The glucoregulatory system of the human body can be viewed as a time-varying dynamical system (Equation 4.1) and is commonly described mathematically using compartmental models. Compartmental models divide the body into several compartments to model characteristics of the overall system [Sanchez, 2019]. The dynamical system states ($\boldsymbol{x}$) represent the state of these compartments (e.g., The UVA/-PADOVA model consists of states associated to glucose intestinal absorption, glucose renal extraction, endogenous glucose production, ... [Kovatchev et al., 2009]), and the control input (u) represents the administered insulin. The system undergoes disturbances ($\boldsymbol{w_d}$) due to activities such as meals and exercise and also features time-varying biological parameters ($\boldsymbol{w_\beta}$).

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = f(\boldsymbol{x}, t, u, \boldsymbol{w_\beta}, \boldsymbol{w_d}). \tag{4.1}$$

Several glucoregulatory models have been proposed in previous work, where the Minimal Model for Glucose Kinetics (1981) [Bergman et al., 1981], Sorenson's model (1985) [Sorensen, 1985], Hovorka's Model [Wilinska et al., 2010], and the UVA/-PADOVA simulator [Kovatchev et al., 2009; Visentin et al., 2018] are some of the widely used models. The recent focus on additional inputs for APS improvement has prompted the development of simulators incorporating supplementary physiological signals and is discussed in detail in Chapter 7. The mGIPsim simulator (2019) [Rashid et al., 2019] has incorporated physiological variables such as Heart Rate (HR), skin temperature, accelerometer, and Energy Expenditure (EE). Resalat et al (2019) [Resalat et al., 2019a] have also incorporated an exercise model in their virtual patient population.

The UVA/PADOVA simulator [Kovatchev et al., 2009] was approved by the FDA as a replacement for animal testing, simplifying the process of pre-clinical proof-of-concept testing. The UVA/PADOVA simulator is currently the only such approved simulator. It comprises adult, adolescent, and children T1D cohorts with ten subjects each. The availability of a cohort rather than a single patient model aids in representing patient variability found in a real T1D population, making meaningful statistical results available for evaluation. The simulator is developed using Matlab [Li, 2001] and can simulate the effects of both insulin and glucagon in open-loop and closed-loop configurations. Different meal scenarios (meal amount, mealtime, and duration)

can be simulated using different commercially available insulin pumps and continuous glucose sensing hardware models [Dalla Man et al., 2014]. It also simulates the errors associated with the selected sensors and pumps. Additional information on the *in-silico* subjects, such as the weight, metabolic tests, and T1D treatment ratios (TDI, Basal Rate), are also provided. The latest version of the simulator can simulate exercise, model the impact of medications, intra-day variability of insulin sensitivity, and the dawn phenomena, and includes updated models of glucagon and insulin delivery [Visentin et al., 2018].

This thesis uses the Simglucose simulator [Xie, 2018], which is an open-source Python implementation of the FDA-approved UVA/PADOVA (2008) model [Kovatchev et al., 2009]. In contrast to other simulators, Simglucose can leverage the existing PyTorch [Paszke et al., 2019] ML framework to develop control algorithms for APS, which allows the integration of RL. Furthermore, Simglucose is freely available compared to others which require the purchase of licenses [Dalla Man et al., 2014; Visentin et al., 2018]. This enables the reproducibility of the work presented in this thesis.

4.2.1.2 Environment Design

In RL, the environment is typically an MDP where all states are known. The environment for the closed-loop glucose control problem in T1D is the glucoregulatory system of the human body, which is a partially observable complex dynamical system [Kovatchev et al., 2009]. The glucoregulatory system is observed through noisy glucose sensor measurements and controlled via an insulin infusion pump. A Partially Observable Markov Decision Process (POMDP) is defined as a tuple $\langle S^\star, S, O, \mathcal{A}, P, R \rangle$, where $S^\star$ is the set of true environment states, S the set of states observed by an observation function O, and $\mathcal{A}$ the set of actions. The transition function P: $(s^\star, a) \rightarrow s'$ represents the system dynamics, where at each step, the RL agent is in a state $s^\star \in S^\star$, takes an action $a \in \mathcal{A}$, and moves from $s^\star$ to the next state $s' \in S^\star$. The observation function $O : s^\star \rightarrow s$ maps the true environment states to the observed states $s \in S$, while the reward function R: $(s, a) \rightarrow r$ provides a reward $r \in \mathbb{R}$ for taking action a at an observed state s. The task of the RL agent is to achieve a defined goal by learning a mapping from states to actions which is called a policy $\big(\pi(a|s)\big)$. The characteristics of the environment are summarised in Table 4.1.

Table 4.1: Summary of environment characteristics.

Dimension	**Characteristic**
Observations	Continuous
Actions	Continuous
Time	Discrete
Dynamics	Stochastic
Observability	Partial
Agency	Single agent
Uncertainty	Uncertain
Termination	Continuing
Stationarity	Non-stationary
Synchronicity	Asynchronous
Delays	Yes
Reality	Simulated

4.2.1.3 State Space Design

The observation function O: $s_t^\star \rightarrow (g_{t-k:t}, i_{t-k:t})$ is designed to map the true states $s_t^\star$ at time t to glucose sensor observation g_t and administered insulin i_t augmented by their past k historical measurements. Hence, the observed state space is defined as,

$$s_t = (g_{t-k:t}, i_{t-k:t}). \tag{4.2}$$

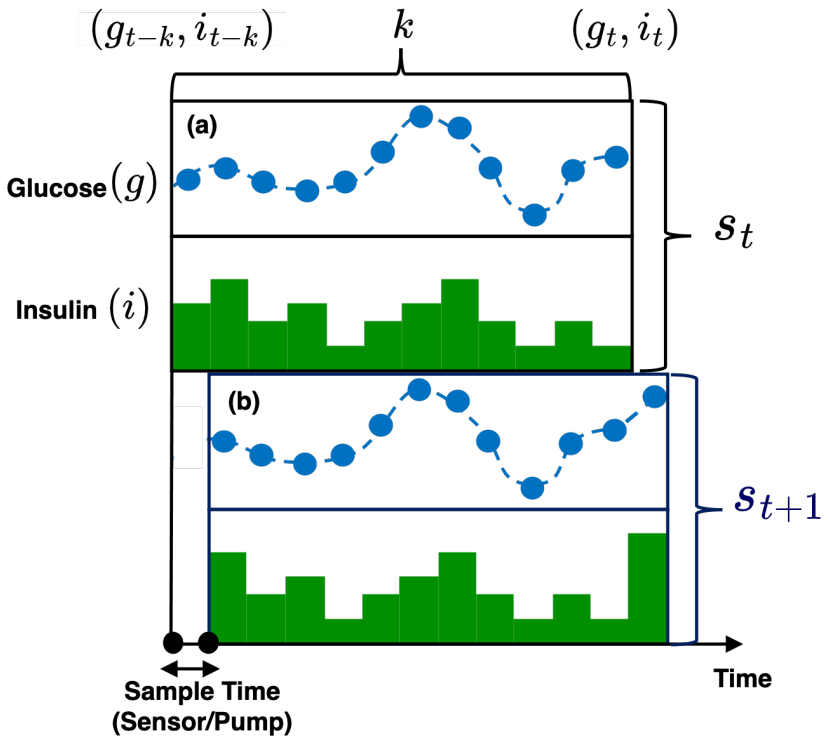


Figure 4.1: State space design. (a) represents the state space at (s_t), while (b) represents the state space at (s_{t+1}). The state space includes glucose (g) (top) and insulin (i) (bottom) measurements based on the sampling frequency of the sensor and pump.

4.2.1.4 Action Space Design

Most previous work focused on only controlling basal insulin and used handcrafted discrete insulin actions. However, this thesis aimed to automate both basal and bolus insulin to design fully-automated APS. Therefore, a continuous action space $\mathcal{A}([-1,1])$, which included both the basal and bolus insulin ranges, was used in this work. The choice of a continuous action space was made to provide additional flexibility to the RL agent to learn a control strategy as opposed to a handcrafted discrete action space. The insulin infusion rate of the insulin pump $I_{pump} \in [0,5]$ U/min was defined as the control space. The action space was mapped to the control space using a non-linear translation function introduced in Chapter 6 according to Equation (4.3), where the parameter η was set to 4.0 and I_{max} to 5 U/min. A thorough analysis of the proposed design and action-space representations are presented in Chapter 6.

$$I_{pump} = I_{max} \cdot e^{\eta(a-1)}, a \in [-1,1]. \tag{4.3}$$

4.2.1.5 Reward Function Design

The reward function is formulated based on the blood glucose Risk Index (RI), similar to the work of Fox et al. [2020]; Hettiarachchi et al. [2022b]. The RI proposed by Kovatchev et al. [2005] weights the risk of blood glucose levels based on the clinical metrics Low Blood Glucose Risk Index (LBGI) and High Blood Glucose Risk Index (HBGI) (Figure 4.2). Additionally, a penalty for hypoglycaemia ($g \leq 39$ mg/dL) is introduced, and the RI normalised ($\overline{RI}$) to [0, -1] for the rest of the glucose range (Equation (4.4)). The penalty is introduced to discourage severe hypoglycaemia, which could result in the death of a person.

$$R(s_t, a_t) = \begin{cases} -15 & g_{t+1} \leq 39 \text{ mg/dL} \\ -1 \cdot \overline{RI}(g_{t+1}) & \text{else} \end{cases}. \tag{4.4}$$

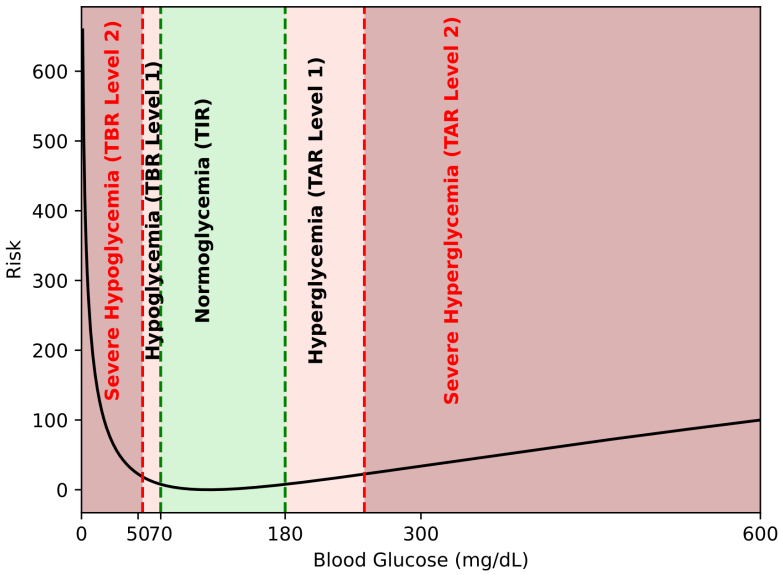


Figure 4.2: Blood glucose risk index.

4.2.2 Implementing State-Of-The-Art RL Algorithms

RL algorithms can be broadly classified as **value-based** and **policy-based** methods. **Actor-critic** is a type of method that lie in the intersection between value- and policy-based methods leveraging the advantages of both. The majority of previous work used value-based RL approaches to design hybrid APS, which focused on only controlling basal insulin formulated as a handcrafted small discrete action space. However, bolus insulin requires a dynamic range to counter meals with different CHO contents, which presents a large continuous action space. This thesis aims to design RL-based APS capable of fully automating insulin infusion (i.e., basal and bolus insulin). **Policy-gradient** is a type of policy-based RL method that offer practical advantages for leveraging large continuous action spaces and learning stochastic policies [Sutton and Barto, 2018]. Hence, this work focuses on policy-gradient methods. Specifically, the focus is on **A2C**, **PPO**, and **SAC** algorithms which are also **actor-critic** methods. Actor-critic methods consist of two networks: (1) an actor network to learn a policy and (2) a critic network to learn the value function. Thus, these methods can be considered as a combination of value-based and policy gradient approaches.

RL tasks can be classified as **episodic tasks**, which can be naturally broken down into identifiable episodes and **continuing tasks**, those which go on continually without limit [Sutton and Barto, 2018]. The glucose control problem can be identified as a continuing task. The RL objective can be formulated using either (1) the standard RL objective or (2) the maximum entropy RL objective. The A2C and PPO algorithms are constructed based on the standard RL objective, while SAC is based on the maximum entropy RL objective. This section first presents the standard RL objective to formulate the A2C and PPO algorithms. Next, the maximum entropy objective is presented to formulate the SAC algorithm.

The standard RL objective aims to maximise the expected sum of the rewards. In continuing tasks, the standard RL objective can be formulated either using the **discounted** or **average reward** RL setting. Discounted RL is a popular approach, where a discount rate $0 \leq \gamma \leq 1$ is introduced to discount the future rewards. Depending on the value γ, the agent could be either myopic or farsighted. Selecting the correct γ is important in this approach [Naik et al., 2019]. In the average reward RL setting, immediate and delayed rewards are equally important to the RL agent. In glucose control, immediate glucose levels are equally important to overall glucose levels as failures in the short-term to manage glucose could lead to catastrophic failures (e.g., short-term severe hypoglycaemia / severe hyperglycaemia). Identifying the optimal discount rate is complicated, as it would depend on individual glucose dynamics. In previous work, Lee et al. [2020b] personalised the discount rate by conducting glucose clamp tests. However, such personalised tuning is challenging in real-world applications. Hence, this thesis focuses on the average reward RL setting, which is more naturally suited for the glucose control problem. Average reward RL can also be identified as a more generalised formulation of the discounted RL problem [Naik et al., 2019; Sutton and Barto, 2018]. A thorough analysis of the discounted and average reward RL is presented in Naik et al. [2019]; Sutton and Barto [2018].

Average Reward Reinforcement Learning Objective. The goals of the average reward RL objective is to maximise the average reward [Naik et al., 2019; Sutton and Barto, 2018] while following a control policy π defined as,

$$R_{avg}(\pi) \doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[r_t | s_0, a_{0:t-1} \sim \pi]. \tag{4.5}$$

The **policy** π induces a value function

$$v^{\pi}(s_t) \doteq \mathbb{E}[G_t | s_t], \tag{4.6}$$

which estimates the expected **return** G_t when starting from state s_t and following the policy π. The return G_t is estimated according to

$$G_t \doteq (r_{t+1} - \hat{R}_t^{avg}) + (r_{t+2} - \hat{R}_t^{avg}) + \cdots + (r_{t+n} - \hat{R}_t^{avg}) + \hat{v}^{\pi}(s_{t+n}), \tag{4.7}$$

where n denotes the number of transition steps, and $\hat{v}^{\pi}(s_{t+n})$ is the bootstrapped **value function** estimate at end state s_{t+n}. $\hat{R}_t^{avg}$ is an estimate of $R_{avg}(\pi)$ at time t. The **advantage function**

$$A^{\pi}(s_t, a_t) \doteq G_t - v^{\pi}(s_t) \tag{4.8}$$

is defined as the difference between the return and the value function estimate and measures whether a target action is better or worse than the average actions. The n step returns were used to compute the advantage function and value function targets ($\hat{v}_t^{target}$). The aim of the critic network (v_ϕ) is to learn the value function to predict the expected return for a given state. The critic network was optimised to minimise the objective function:

$$L^{v}(\phi) = \hat{\mathbb{E}}_t \left[\frac{1}{2} \left(v_\phi(s_t) - \hat{v}_t^{target} \right)^2 \right]. \tag{4.9}$$

4.2.2.1 Advantage Actor Critic (A2C)

In policy-gradient methods, a parameterised policy ($\pi(a|s, \theta)$ is optimised based on the gradient of a scalar performance measure according to the policy gradient theorem to learn an optimal policy [Sutton and Barto, 2018]. In A2C and PPO algorithms the performance measure is based on the advantage function. The n step returns were used to compute the advantage function targets ($\hat{A}_t$). The actor network objective for A2C is to maximise the objective function ($L_{A2C}^{\pi}(\theta)$) defined in Equation (4.10). The entropy of the policy ($H(\pi(\cdot|s_t))$) is added to the objective to facilitate exploration weighted by a hyperparameter β_s.

$$L_{A2C}^{\pi}(\theta) = \hat{\mathbb{E}}_t \left[log \pi_\theta(a_t|s_t) \hat{A}_t + \beta_s H\Big(\pi(\cdot|s_t) \Big) \right]. \tag{4.10}$$

4.2.2.2 Proximal Policy Optimisation (PPO)

The PPO algorithm [Schulman et al., 2017] imposes constraints on policy updates to avoid excessive changes between the old policy ($\pi_{\theta_{old}}$) and the new policy (π_θ), which is valuable in safety critical applications, where excessive changes in the control strategy could lead to unexpected behaviour. This is achieved by clipping the probability ratios of the new and old policies at $1-\epsilon$ or $1+\epsilon$ as shown in Equation (4.11). The constrained optimisation also improves the sample efficiency of the algorithm, by enabling multiple updates for a given data sample. The actor network objective for PPO is to maximise the objective function ($L^{\pi}_{PPO}(\theta)$).

$$L^{\pi}_{PPO}(\theta) = \hat{\mathbb{E}}_t\left[min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t, clip\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right) + \beta_s H\Big(\pi(\cdot|s_t)\Big)\right]. \tag{4.11}$$

4.2.2.3 Soft Actor Critic (SAC)

Maximum Entropy Reinforcement Learning Objective. The SAC algorithm [Haarnoja et al., 2018] is an off-policy algorithm designed based on the maximum entropy objective which generalises the standard RL objective by augmenting it with an entropy term (Equation (4.12)). The goal is to learn an optimal policy which also maximises its entropy at each visited state. The temperature parameter α determines the relative importance between the entropy term and the reward.

$$R_{entropy}(\pi) \doteq \sum_{t=1}^{\infty}\mathbb{E}\left[\Big(r_t + \alpha H\Big(\pi(\cdot|s_t)\Big)|s_0, a_{0:t-1} \sim \pi\right]. \tag{4.12}$$

The action-value function or **Q-function** is defined as the value of taking action a_t in state s_t under a policy π and denoted as $q^\pi(s_t, a_t)$ (Equation (4.13)).

$$q^\pi(s_t, a_t) \doteq \mathbb{E}[G_t|s_t, a_t \sim \pi]. \tag{4.13}$$

The return G_t is defined based on the discounted RL setting,

$$G_t \doteq r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^{t+n-1} r_{t+n-1} + \gamma^{t+n}\hat{v}^\pi(s_{t+n}). \tag{4.14}$$

Unlike A2C and PPO, in SAC the critic network aims to learn the soft Q-function. The critic is implemented using two soft Q-functions (Q_{ϕ_j}) and two target Q-functions ($Q_{\bar{\phi}_j}$) to mitigate the positive bias. The n step returns were used to compute the action-value function targets ($\hat{Q}^{target}_t$). The soft Q-function is learnt through minimising the objective function ($L^{Q}_{SAC}(\phi_j)$) in Equation (4.15). The parameters of the target Q functions are updated using soft replacement as shown in Equation (4.16).

$$L^{Q}_{SAC}(\phi_j) = \hat{\mathbb{E}}_t \left[\frac{1}{2} \left(Q_{\phi_j}(s_t) - \hat{Q}^{target}_t \right)^2 \right] \text{ for } j \in \{1,2\}. \tag{4.15}$$

$$\bar{\phi}_j = \tau\phi_j + (1-\tau)\bar{\phi}_j \text{ for } j \in \{1,2\}. \tag{4.16}$$

The objective of the actor network is to update the policy towards the exponential of the new soft Q function and the resulting optimisation objective is presented in Equation (4.17).

$$L^{\pi}_{SAC}(\theta) = \hat{\mathbb{E}}_t \Big[\alpha log(\pi_\theta(a_t|s_t)) - Q_\phi(s_t, a_t) \Big]. \tag{4.17}$$

The parameter α is learned automatically by minimising the objective function in Equation (4.18). This ensures that the policy explores more the regions where the optimal actions are uncertain and remain more deterministic in other regions. $\bar{H}$ is a hyperparameter representing the target entropy.

$$L^{\alpha}_{SAC}(\alpha) = \hat{\mathbb{E}}_t \Big[-\alpha log\pi_t(a_t|s_t) - \alpha\bar{H} \Big]. \tag{4.18}$$

4.2.2.4 Design and Implementation Details

The neural network architectures of the three candidate RL algorithms are similar in design. The actor and critic networks were implemented as two separate networks and consisted of a feature extractor module implemented using a single-layer LSTM network with 16 hidden units. The policy, value, and Q-function modules were implemented using 3 dense layers with 32 units each and connected to the hidden state of the LSTM network. The networks were implemented using PyTorch [Paszke et al., 2019] and optimised using the Adam optimiser [Kingma and Ba, 2014]. ReLU activation functions were used in the two networks.

In A2C and PPO the policy π was implemented as a Gaussian distribution where tanh and sigmoid activation functions were used to predict the mean and standard deviation of the policy. SAC used a similar structure. However the gaussian policy implemented was unbounded and an invertible squashing function was used as presented in the original work of Haarnoja et al. [2018]. The hyperparameters of the algorithms are presented in Table 4.2.

4.2.3 Clinical Benchmarks

The candidate RL algorithms were benchmarked against a basal-bolus (BB) clinical insulin treatment strategy with meal announcement. For BB, a basal insulin infusion rate is used to provide background insulin requirements, while a meal insulin bolus dose is used to counter meals and a correction insulin bolus dose to counter high blood glucose levels. BB treatment was used as the gold standard benchmark

Table 4.2: Hyperparameters for A2C, PPO, and SAC.

Hyperparameter	**Value**
Shared (A2C, PPO, SAC)	
Sample time	5 minutes
Glucose sensor	Guardian RT
Insulin pump	Insulet
Augmented state history	12 (1-hour)
Total number of Interactions	800,000
Optimiser	Adam [Kingma and Ba, 2014]
Number of steps per rollout (n)	256
Number of workers	16
Learning rate of policy/value	3 x 10^{-4}
PPO	
Batch size of policy/value	1,024
No. of policy/value epochs	5
Entropy Coefficient (β_s)	0.001
PPO clip range (ϵ)	0.1
Target Kullback-Leiber divergence threshold	0.01
SAC	
Replay buffer size	100,000
Discount rate (γ)	0.997
Batch size	256
Target smoothing coefficient (τ)	0.005
Target update interval	1
Gradient steps	1
Initial entropy coefficient	0.1
Target entropy ($\bar{H}$)	-1

Acronyms: A2C - Advantage Actor Critic, PPO - Proximal Policy Optimisation, SAC - Soft Actor Critic.

for evaluation due to its widespread recognition among clinicians [Bergenstal et al., 2010]. This treatment strategy is based on patient-specific characteristics and requires prior knowledge about the CHO content of future meals (typically 20 minutes in advance).

Two versions of the BB treatment were replicated based on the previous work of Fox et al. [2020]; the Basal-Bolus Ideal (BBI) case, where the CHO of announced meals was provided accurately, and the realistic case, which considered the human inaccuracy in CHO estimation named Basal Bolus Human Error (BBHE). The CHO counting error was calculated using a mathematical model developed to conduct *in-silico* trials [Roversi et al., 2020]. The model used the real CHO content of the meal and meal type to simulate the decision of the patient with T1D.

For the replication of the two BB methods, the patient-specific characteristics provided by the T1D simulator and parameters proposed in previous work [Fox et al., 2020] were used. The final insulin delivered over time I_t is presented in Equa-

tion (4.19). Here c_t represents the meal CHO estimate, g_t the current glucose value, and g_{target} (140 mg/dL) the target glucose for correction boli. The Total Daily Insulin (TDI) provided by the simulator for each subject was used to personalise the Basal Rate $\left(\text{BR} = 0.48 \cdot \text{TDI U/day}\right)$, Carbohydrate Insulin Ratio $\left(\text{CIR} = \frac{500}{\text{TDI}} \text{ g/U}\right)$, and Insulin Sensitivity Factor $\left(\text{ISF} = \frac{1800}{\text{TDI}} \text{ mg/dL per U}\right)$ [Walsh et al., 2011]. The meal bolus was calculated as $\frac{c_t}{\text{CIR}}$ U, and delivered 20 minutes before a meal.

The correction insulin bolus was calculated as $\frac{(g_t - g_{target})}{\text{ISF}}$ U, and was delivered when the glucose level increased above the correction threshold of 150 mg/dL. The correction insulin dose was replicated according to Fox et al. [2020] and was only applied during meal events ($c_t > 0$). The application of the correction bolus had a further condition, where it was only applied when no other meals were present for a past 3-hour duration. This ensured that each meal was only corrected once. The *cool* parameter was set to match this criterion, where *cool* was set to 1 if no other meals were present for the 3-hour duration and to 0 otherwise.

$$I_t = \text{BR} + (c_t > 0) \cdot \left(\frac{c_t}{\text{CIR}} + cool \cdot \frac{g_t - g_{target}}{\text{ISF}} \right). \tag{4.19}$$

4.2.4 Evaluation Metrics

Both clinical metrics and RL-based metrics were used for the evaluation. The clinical metrics included the standard T1D metrics of glucose risk indices (RI, HGBI, LGBI) [Kovatchev et al., 2005] and the percentage of time spent in different glucose regions. The definitions of the clinical metrics were presented in Table 2.1. The clinical objective was to minimise all risk indices, improve the time spent in the normoglycaemic range (TIR), and minimise the time spent in other glucose regions (severe hypoglycaemia, hypoglycaemia, hyperglycaemia, and severe hyperglycaemia). Additionally, ***catastrophic failures*** were defined as those simulations that recorded glucose levels outside the detectable range (40–600 mg/dL) of the glucose sensor[1]. Such simulations were terminated, and a Failure Rate (FR) was calculated as the number of catastrophic failures over the total evaluation simulations.

As the RL metric, the total reward achieved by an RL algorithm in evaluation simulation ($R^{eval} \in [-15, 288]$) was used. The theoretical maximum total reward achievable by following a perfect glucose control strategy was bounded by $R^{eval}_{max} = 288$. The total reward achieved by an RL algorithm in an evaluation simulation (R^{eval}), as a percentage of the maximum achievable reward ($R^{\star}$), was defined as the Percentage Reward ($PR = (\frac{R^{eval}}{R^{\star}}) * 100$). The RL objective was to maximise the PR metric.

[1] A glucose level $\leq$ 40 mg/dL is life-threatening and can result in major cardiovascular and cerebrovascular problems [Kalra et al., 2013]. A glucose value $\geq$ 600 mg/dL is also a life-threatening emergency referred to as the Hyperosmolar hyperglycaemic state [Stoner, 2005].

4.2.5 Experiments

The candidate RL algorithms were trained *in-silico* using the 10 adult and 10 adolescent cohorts of the Simglucose T1D simulator [Xie, 2018]. The Insulet pump and the GuardianRT glucose sensor provided by the simulator were used with a sampling time of 5 minutes and a meal protocol was implemented to conduct simulations [Hettiarachchi et al., 2022b].

This thesis defines two meal protocols; a **training protocol** for training the RL-based algorithms and an **evaluation protocol** to evaluate treatment strategies. For the training, a challenging yet realistic meal protocol where meals were uniformly randomised based on CHO content and time was designed (Table 4.3). The randomisation of the meal content and occurrence increased the uncertainty, and the large CHO meals increased the disturbance to glucose control. The probability of occurrence of a meal was set to 0.95 to replicate missed meals by the user. The evaluation meal protocol used in Lee et al. [2020b] was chosen for all algorithms to support comparisons with other works. The evaluation meal protocol spanned 24 hours starting at 00:00hrs fixed with three meals: 40g of CHO for breakfast at 8:00hrs, 80g of CHO for lunch at 13:00hrs, and 60g of CHO for dinner at 20:00hrs.

Table 4.3: Training meal protocol.

Meal Type	**Time** [*hours*]	**Carbohydrate (CHO) content** [g]
Breakfast	07:00–09:00	30–60
Lunch	12:00–14:00	70–100
Dinner	19:00–21:00	50–110

The RL algorithms were trained for multiple iterations until $800,000$ interactions were completed, and separate algorithms were trained for each *in-silico* subject. An *interaction* was defined as an insulin delivery action taken every 5 minutes. For each subject, the experiments were conducted for three different random seeds. The training of the candidate RL algorithms alternated between sampling and optimisation phases conducted for multiple iterations. After the conclusion of each training iteration, 60 evaluation simulations (i.e., 3 random seeds x 20 simulations/iteration) were conducted for each subject. These simulations provided insights into the training characteristics of the RL algorithms. Once training was fully complete (at $800,000$ interactions), the RL algorithms were evaluated using $1,500$ evaluation simulations (i.e., 3 random seeds x 500 simulations/subject) conducted for each subject. Hence, the evaluation protocol was used to analyse the performance of RL-based algorithms during training and to compare all candidate algorithms upon the conclusion of training. The evaluation of candidate RL algorithms and benchmark clinical treatment methods was performed using the same two cohorts of the simulator. The initial blood glucose level was restricted between 110–130 mg/dL for the evaluation simulations to ensure fair comparisons between the methods. The BBI and BBHE approaches were also evaluated based on $1,500$ evaluation simulations conducted for each subject for comparison.

4.3 Results

RL Analysis. The PPO algorithm performed best compared to A2C and SAC in terms of the final reward performance on both adult and adolescent cohorts achieving a PR of 91.54% and 86.53% respectively (Table 4.4). The comparatively lower PR on the adolescent cohort highlights the challenging nature of glucose control in adolescents. The training curves of the *in-silico* subjects provide insight into the learning behaviour of the RL algorithms (Figure 4.3, 4.4). Stable learning was observed for PPO across all subjects and was able to outperform A2C and SAC for all 20 *in-silico* subjects while also being the most efficient.

The SAC algorithm performed worst and failed to converge in some subjects (e.g., Adult7, Adolescent4, Adolescent 5), while an unstable learning behaviour was observed in almost all subjects, where large fluctuations in total reward were observed (Figure 4.3, 4.4). This could be due to the off-policy nature of the algorithm. The SAC algorithm also recorded a much higher FR compared to the PPO and A2C algorithms for both cohorts, where the FR was 59.49% and 82.06% for the adult and adolescent cohorts respectively.

A2C performed well compared to SAC. However, the efficiency in learning was low compared to PPO, where a larger number of steps were required for convergence (e.g., Adult0, Adolescent0) (Figure 4.3, 4.4). This could be due to the enhanced sample efficiency of PPO. The FR in A2C was also higher compared to PPO, which was 9.11% and 14.41% compared to the 2.79% and 4.93% in PPO for the adult and adolescent cohorts respectively.

Table 4.4: Performance comparison of candidate glucose control algorithms averaged for adult and adolescent cohorts. Detailed clinical results are presented in Table 4.5, 4.6.

Algorithm	TIR (%)	Failure Rate (%)	Percentage Reward (%)
Adults			
BBI	71.02±11.29	0.39	-
BBHE	69.78±11.29	0.35	-
A2C	59.06±14.31	9.11	84.86±16.23
PPO	69.12±10.53	2.79	91.54±8.88
SAC	61.76±21.01	59.49	50.71±33.97
Adolescents			
BBI	71.43±12.31	0.00	-
BBHE	70.23± 12.52	0.00	-
A2C	56.03±14.40	14.41	78.70±16.51
PPO	63.72±13.95	4.93	86.53±10.67
SAC	65.62±20.16	82.06	37.22±30.82

Acronyms: A2C - Advantage Actor Critic, BBHE - Basal Bolus Human Error, BBI - Basal Bolus Ideal, PPO - Proximal Policy Optimisation, SAC - Soft Actor Critic, TIR - Time In Range.

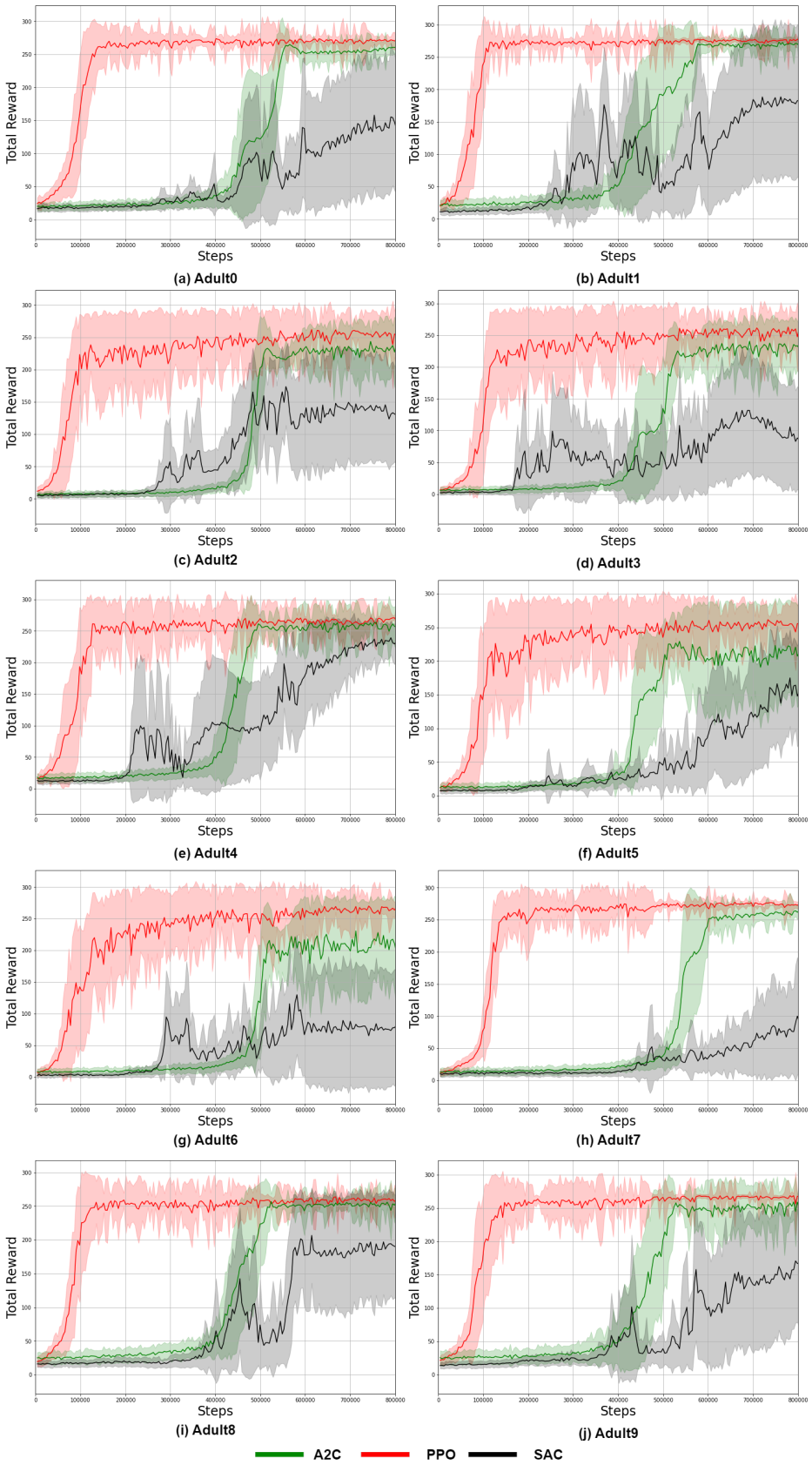


Figure 4.3: Comparison of A2C, PPO, SAC algorithms during training on the adult cohort. The mean and standard deviation (shaded area) of the total cumulative reward for evaluation simulations (3 random seeds x 20 evaluation scenarios / iteration) achieved is presented against 800,000 learning interactions.

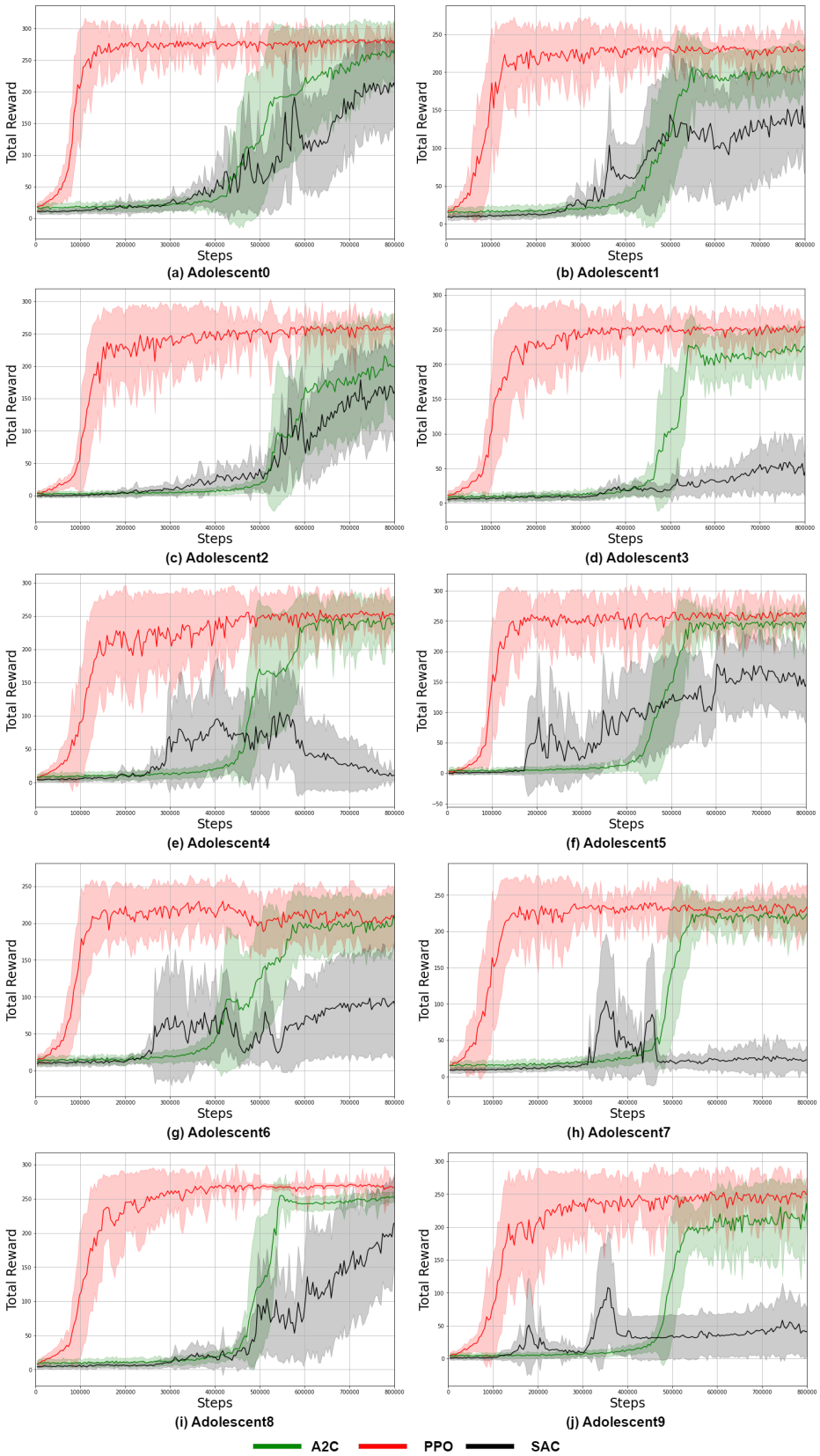


Figure 4.4: Comparison of A2C, PPO, SAC algorithms during training on the adolescent cohort. The mean and standard deviation (shaded area) of the total cumulative reward for evaluation simulations (3 random seeds x 20 evaluation scenarios / iteration) achieved is presented against 800,000 learning interactions.

Clinical Analysis. As expected, reflecting the reward performance, PPO performed better in terms of the clinical metrics of TIR and failures compared to A2C and SAC (Table 4.4). BBI and BBHE methods based on basal-bolus clinical treatment were used as the gold standard benchmark. It is essential to highlight that these methods use meal announcements 20 minutes in advance to mitigate the disturbances of meal events and CHO estimation. Hence, the direct comparison of BBI and BBHE is very challenging for the autonomous RL-based methods explored in this thesis. Thus, there is a trade-off between the control performance and the cognitive burden on people with T1D. Therefore, this thesis, aims to achieve the established gold standard performance while eliminating the cognitive burden.

The TIR performance of PPO (69.1%) was slightly below BBI (71.0%) and BBHE (69.8%) approaches in the adult cohort, while it was lower in the challenging adolescent cohort, where PPO only achieved 63.7% compared to 71.4% and 70.2% in the BBI and BBHE respectively. The failure rate was also higher in PPO compared to the BB approaches. As expected, the TIR performance of BBHE was lower than BBI due to the replicated human error in BBHE. A detailed comparison of clinical metrics between the candidate algorithms is presented in Table 4.5 and 4.6.

The insulin administration strategies learnt by the RL-based algorithms were much more complex compared to the BB strategies (Figure 4.5). This highlights the importance of focussing on the explainability of RL-based APS. Example simulated glucose control trajectories for the candidate algorithms focusing on Adolescent0 are presented in Figure 4.5 (in Chapter 8, an online demonstration tool is introduced, which can be used to run custom simulations for the *in-silico* T1D subjects to visualise the performance of the candidate algorithms).

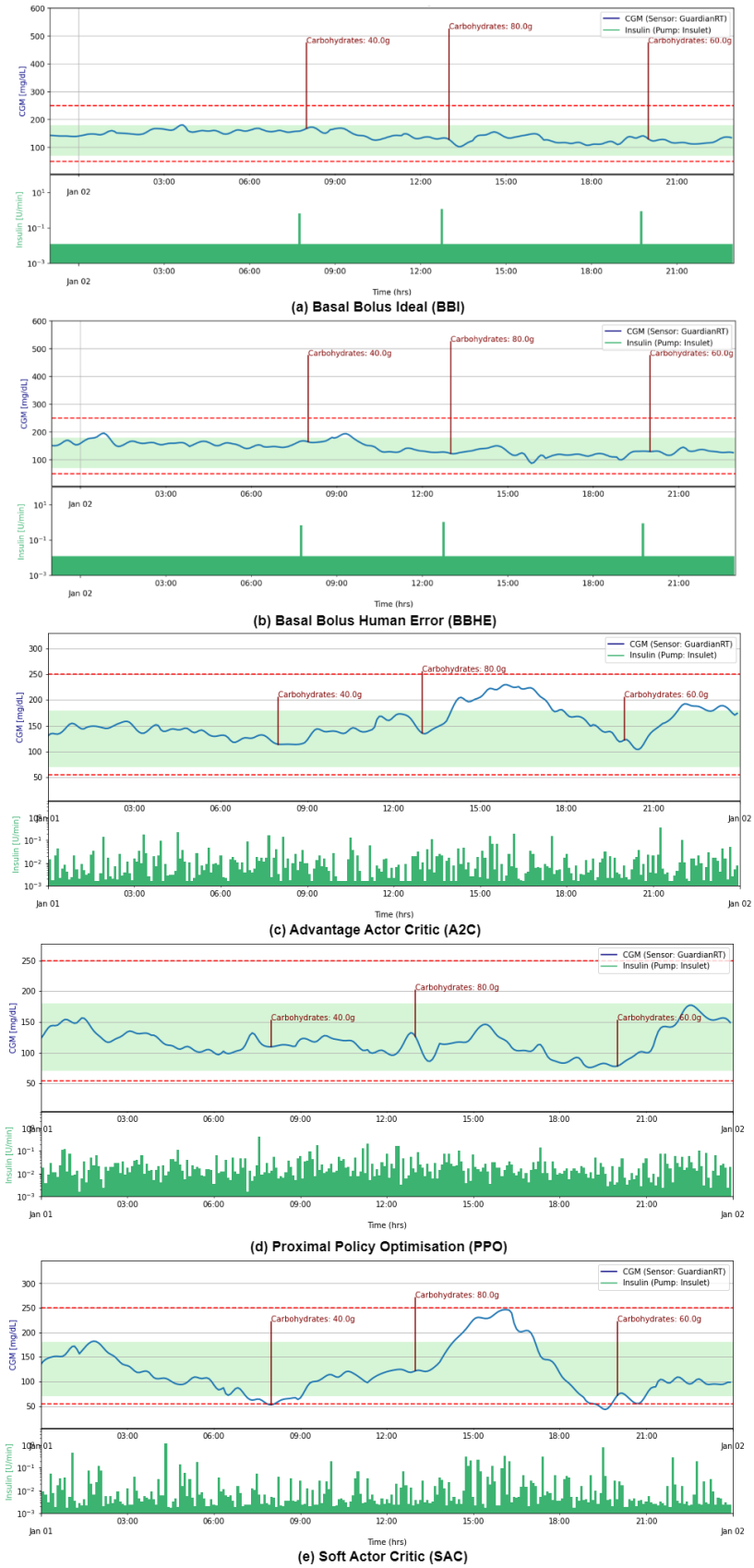


Figure 4.5: Sample glucose simulation trajectories for BBI, BBHE, A2C, PPO, and SAC based on Adolescent0.

Table 4.5: Comparison of clinical performance for the state-of-the-art RL algorithms and basal-bolus treatment for the adult cohort.

Method	Failure (%)	Severe Hypo.(%) TBR-Level 2	Hypo. (%) TBR-Level 1	Normo. (%) TIR	Hyper. (%) TBR-Level 1	Severe Hyper.(%) TBR-Level 2	RI	LBGI	HBGI
BBI	0.39	0.00†	0.00	70.83	22.57	2.08	7.68	0.81	6.72
		0.00-0.00*	0.00-0.00	61.46-79.17	17.36-31.94	0.00-4.86	6.07-9.38	0.28-1.68	4.97-7.94
		0.06(0.40)‡	0.64(2.01)	71.02(11.29)	24.44(10.57)	3.85(5.72)	8.35(4.05)	1.33(1.60)	7.02(2.79)
BBHE	0.35	0.00	0.00	69.79	23.61	3.12	7.62	0.41	6.92
		0.00-0.00	0.00-0.00	60.42-78.47	18.06-32.29	0.00-5.21	5.78-8.69	0.11-0.93	5.11-8.11
		0.05(0.36)	0.40(1.42)	69.78(11.29)	25.44(10.57)	4.33(5.64)	8.00(3.74)	0.88(1.37)	7.11(2.67)
A2C	9.11	0.00	0.00	58.68	23.61	14.24	12.07	0.54	10.50
		0.00-0.00	0.00-1.04	48.82-68.75	17.36-29.51	6.25-22.92	9.15-15.90	0.03-2.64	7.41-13.95
		0.84(2.93)	1.48(3.56)	59.06(14.31)	23.12(9.95)	15.49(11.95)	13.15(5.77)	2.14(3.43)	11.00(5.44)
PPO	2.79	0.00	0.00	69.44	20.83	7.99	9.56	0.89	8.05
		0.00-0.00	0.00-1.04	62.15-76.04	16.32-25.35	2.43-12.85	6.89-12.04	0.28-2.17	5.86-10.10
		0.32(1.51)	1.19(2.77)	69.12(10.53)	20.93(7.11)	8.44(6.88)	9.79(3.66)	1.64(2.11)	8.14(3.05)
SAC	59.49	2.64	2.86	66.67	12.85	1.77	14.69	7.25	6.01
		0.00-8.45	0.00-7.83	45.83-78.95	0.00-22.92	0.00-27.08	10.90-21.82	1.27-12.07	1.44-16.06
		5.31(6.86)	5.11(6.17)	61.76(21.01)	13.13(12.25)	14.69(19.70)	17.39(9.11)	7.27(5.69)	10.12(10.55)

The †median, *inter-quartile range followed by the ‡mean (standard deviation) are presented. Acronyms: A2C - Advantage Actor Critic, BBHE - Basal Bolus Human Error, BBI - Basal Bolus Ideal, HBGI - High Blood Glucose Index, LBGI - Low Blood Glucose Index, PPO - Proximal Policy Optimisation, RI - Risk Index, RL - Reinforcement Learning, SAC - Soft Actor Critic, TAR - Time Above Range, TBR - Time Below Range, TIR - Time In Range.

Table 4.6: Comparison of clinical performance for the state-of-the-art RL algorithms and basal-bolus treatment for the adolescent cohort.

Method	Failure (%)	Severe Hypo.(%) TBR-Level 2	Hypo. (%) TBR-Level 1	Normo. (%) TIR	Hyper. (%) TBR-Level 1	Severe Hyper.(%) TBR-Level 2	RI	LBGI	HBGI
BBI	0.00	0.00†	0.00	67.71	16.32	4.86	7.93	0.43	7.56
		0.00-0.00*	0.00-0.00	63.54-76.39	11.11-27.43	0.00-18.40	5.21-14.10	0.09-1.40	5.04-12.26
		0.03(0.27)‡	0.44(1.27)	71.43(12.31)	19.65(11.53)	8.47(8.82)	9.26(4.83)	1.07(1.57)	8.19(3.78)
BBHE	0.00	0.00	0.00	66.67	18.06	5.56	8.06	0.21	7.94
		0.00-0.00	0.00-0.00	63.19-73.26	11.11-28.82	1.74-18.75	5.68-14.11	0.01-0.87	5.67-12.59
		0.01(0.10)	0.20(0.76)	70.23(12.52)	20.37(12.04)	9.20(8.85)	9.15(4.51)	0.69(1.11)	8.46(3.82)
A2C	14.41	0.00	0.00	54.86	16.32	26.04	17.55	0.49	16.07
		0.00-0.00	0.00-0.69	46.88-62.38	11.46-22.22	17.36-34.72	13.33-23.22	0.01-2.32	11.90-21.67
		0.83(3.04)	1.55(3.84)	56.03(14.40)	16.57(8.30)	25.03(14.35)	18.03(7.12)	1.94(3.16)	16.09(7.60)
PPO	4.93	0.00	0.00	60.42	15.62	17.71	14.66	1.21	12.75
		0.00-0.00	0.00-1.39	54.17-70.14	12.15-20.83	11.11-27.43	11.16-20.63	0.49-2.45	9.73-18.84
		0.28(1.36)	1.30(2.78)	63.72(13.95)	16.15(7.23)	18.55(11.09)	15.40(6.67)	1.82(1.99)	13.58(6.55)
SAC	82.06	4.70	4.94	71.43	3.12	0.00	14.83	9.13	3.06
		1.01-11.54	1.04-9.38	56.25-80.25	0.00-17.50	0.00-17.78	11.62-21.47	4.19-12.62	1.01-13.85
		6.85(7.00)	6.26(6.15)	65.62(20.16)	9.99(12.56)	11.27(16.79)	16.95(7.49)	8.46(5.35)	8.49(9.78)

The †median, *inter-quartile range followed by the ‡mean (standard deviation) are presented. Acronyms: A2C - Advantage Actor Critic, BBHE - Basal Bolus Human Error, BBI - Basal Bolus Ideal, HBGI - High Blood Glucose Index, LBGI - Low Blood Glucose Index, PPO - Proximal Policy Optimisation, RI - Risk Index, RL - Reinforcement Learning, SAC - Soft Actor Critic, TAR - Time Above Range, TBR - Time Below Range, TIR - Time In Range.

4.4 Discussion

In this chapter, first, the glucose control problem was modelled as a POMDP and state/action space, and the reward function was designed to formulate it as an RL task. Next, three state-of-the-art RL algorithms (A2C, PPO, and SAC) were selected for comparisons and benchmarked against standard BB clinical treatment. The PPO algorithm demonstrated superior performance, efficiency, and smooth learning compared to A2C and SAC algorithms. The comparative superior performance of PPO over the other candidate RL algorithms can be attributed to the on-policy constrained policy optimisation procedure, which improves sample efficiency compared to A2C and stability compared to SAC. PPO showed promise in eliminating CHO estimation and meal announcement. Hence, PPO can be shortlisted as a suitable algorithm to explore the development of RL-based APS. Even though the PPO algorithm has potential to reduce the cognitive burden compared to the standard BB clinical treatment approaches, the performance of PPO was lower. Therefore, PPO requires further performance improvements. Specifically, the TIR performance can be improved, and catastrophic failures must be reduced. In the next chapter, a novel RL algorithm named G2P2C is developed based on PPO and its identified favourable characteristics to address these challenges.

The best practice in control algorithm development for APS is first to use FDA-approved simulators Kovatchev et al. [2009], which is also essential in the context of RL-based systems due to safety constraints and the requirement of large amounts of interactions restricting *in-vivo* training. Then as a next step, periodical personalisation of treatment is carried out by clinicians [Nathan et al., 1993]. Hence, designing general algorithms during *in-silico* training that can be personalised during *in-vivo* translation is essential. The proposed problem formulation does not use personalised measures of the *in-silico* subjects compared to previous work [Fox et al., 2020; Lee et al., 2020b]. Lee et al. [2020b] tunes a discount rate based on personalised normoglycaemic glucose clamp tests, and Fox et al. [2020] personalises the action space based on individual basal rates. The average reward RL setting used in this work eliminates the need for a discount rate, and the proposed action space formulation does not require personalisation (a detailed analysis of the proposed action space formulation is presented in Chapter 6). Hence, the proposed generalised formulation is expected to be valuable in translating the research to the real world. However, it is essential to highlight that a data-driven personalisation approach would be valuable for APS (e.g., using a learned glucose dynamics model of an individual for glucose control). This aspect will be explored in the next chapter. Also, it is important to highlight that RL algorithms require significant compute during training and parameter tuning, which makes personalisation for each individual *in-silico* subject highly inefficient and unscalable.

Previous work on RL-based APS mainly focused on Q-learning and DQN-based RL algorithms to design hybrid systems, where the agent was provided handcrafted discrete insulin actions only to control basal insulin [Ngo et al., 2018a,b; Zhu et al., 2019; Fox and Wiens, 2019; Zhu et al., 2020]. Recent advancements in RL algorithms

have focused on continuous control through continuous action spaces (Section 3.4). Hence, the state-of-the-art A2C, PPO, and SAC algorithms were selected in this analysis to control both basal and bolus insulin. PPO and SAC algorithms have also been explored in recent work on RL-based APS [Fox et al., 2020; Lee et al., 2020b; Lim et al., 2021]. A comparison of the RL-based APS proposed in this thesis against previous work is provided in the next chapter. In this study, even though similar neural network architectures and learning parameters were used (Section 4.2.2.4), the performance of SAC was inferior to A2C and PPO algorithms. The stability during training was also inferior in SAC (Figure 4.3, 4.4), which could be due to the fact that it is an off-policy algorithm compared to A2C and PPO, which are on-policy algorithms (Section 3.1). The SAC algorithm is based on a maximum entropy RL objective and was designed using a discounted RL setting. In contrast, the A2C and PPO algorithms were designed using an average reward RL objective.

In popular RL problems and benchmarks, the problem formulations are well established. Hence, they do not require effort to explore different state/action spaces and reward function formulations. However, this task required the exploration of different designs. Moreover, RL algorithms are often brittle to hyperparameter settings (e.g., learning rate, batch size, entropy coefficient). In established benchmarks, stable algorithm implementations and hyperparameter settings are often publicly available and presented in the literature. However, in the problem of glucose control, such implementations and algorithm settings are not available due to the less exploration of the problem. Hence, to facilitate the research in this domain and for reproducibility, this work's implementations and hyperparameter settings are published on GitHub (https://github.com/chirathyh/G2P2C) as open-source under the MIT license.

Chapter 5

G2P2C: A Novel Modular RL Algorithm for Glucose Control

This chapter introduces a novel RL-based algorithm to address the shortcomings of the state-of-the-art algorithms explored in the previous chapter. The proposed algorithm extends the PPO algorithm by introducing two novel optimisation phases, model learning and planning. In view of these novel added features, the algorithm was named G2P2C — Glucose Control by Glucose Prediction and Planning. This chapter presents the algorithm design and implementation details of G2P2C and evaluates its performance against clinical treatment benchmarks and the best-performing state-of-the-art PPO algorithm from the previous chapter.

5.1 Introduction

In the previous chapter, the state-of-the-art RL algorithms were explored to design RL-based APS while addressing the challenges of partial observability, high-dimensional state/action spaces, large sensor/actuator delays, and reward function design through the proposed problem formulation. However, the performance of the proposed system can be further improved compared to basal-bolus clinical treatment, and challenges associated with safety, explainability, and transferability remain. This chapter investigates improvements to APS safety by focussing on practical limitations of the standard RL optimisation objective, which arise from the lack of knowledge to formulate an ideal reward function.

Most real-life control problems often have short-term and long-term objectives, and formulating reward functions to capture both objectives is challenging or even infeasible [Khorasgani et al., 2019]. To focus on the two objectives, model-based RL approaches which leverage short/long-term prediction models have been proposed Khorasgani et al. [2019], while several studies have focused on learning value functions over multiple time horizons [Sutton, 1995; Fedus et al., 2019; Romoff et al., 2019].

The glucose regulation task also has both long- and short-term objectives. The long-term objective is to improve the time spent in the desirable normoglycaemic range, whereas short-term hypoglycaemia and hyperglycaemia are detrimental to

health [DiMeglio et al., 2018]. Ideally, optimising for the long-term through an infinite horizon RL objective is expected to result in learning a policy that is also optimal in the short term. However, in the previous chapter, it was observed that the learnt policies for glucose control are sub-optimal in the short term due to the presence of catastrophic failures. Even though infrequent, these failures are unacceptable for a high-risk medical application since they could result in death (e.g., severe hypoglycaemia in the short term).

APS are categorised as high-risk medical devices, and the best practice for developing control algorithms is first to use FDA-approved simulators Kovatchev et al. [2009]. These algorithms are later fine-tuned and personalised periodically to individuals [Nathan et al., 1993]. Designing RL-based APS, which learns online on real-world people, is practically infeasible as RL algorithms typically require large amounts of experience to learn a stable policy for which years of training is required (At the rate of 5 minutes per learning experience adopted in Chapter 4, training would require more than 9.5 years). Offline strategies would still require extensive experience, and current clinical treatment strategies used to collect such experience may not provide valuable information. This is due to their static rule-based nature [Nathan et al., 1993], where a fixed basal rate is used together with sparse bolus insulin doses, which limits the exploration required in the action space for RL agents. Hence, using simulators as a first step to train RL-based APS is a more suitable approach, while personalising and fine-tuning can be carried out as the next step. It is essential to learn a safe control policy during *in-silico* initial training and focus on safe personalisation during *in-vivo* fine-tuning. This chapter focuses on learning a safe control policy *in-silico*, while research on strategies to safely transfer and personalise the learnt policy to the real world is reserved for future work.

A novel modular RL algorithm was designed to address safety and improve the performance of APS. The proposed algorithm extends PPO with two additional optimisation phases, model learning and planning. In view of these added features, the algorithm was named **G2P2C — Glucose Control by Glucose Prediction and Planning** (Figure 5.1). In the model learning phase, a glucose dynamics model is learned as an auxiliary learning task, followed by the planning phase to fine-tune the learned control policy to a short-time horizon by conducting model-based simulations to improve safety. Therefore the proposed approach could combine the characteristics of both model-free and model-based methods. Furthermore, G2P2C improves sample efficiency and possesses desirable characteristics towards offline personalisation, which is expected to be valuable. The performance of G2P2C was assessed *in-silico* and comparatively evaluated against the state-of-the-art PPO algorithm and basal-bolus clinical treatment benchmarks.

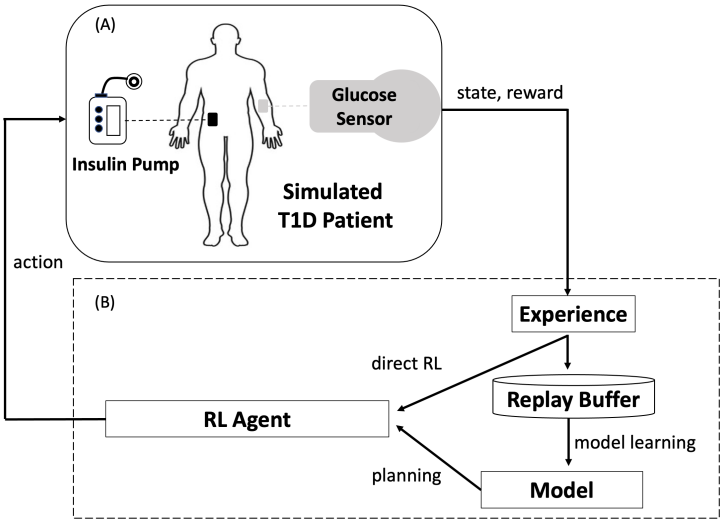


Figure 5.1: G2P2C - Glucose Control by Glucose Prediction and Planning. (A) a glucose sensor and an insulin pump attached to a Type 1 Diabetes (T1D) patient, and (B) the proposed Reinforcement Learning (RL)-based control algorithm G2P2C. G2P2C uses its experience to learn a control policy (direct RL) while storing its experience in a replay buffer. It uses stored data to learn a glucose dynamics model (model learning), which is then used to improve the policy (planning).

5.2 Methodology: Glucose Control by Glucose Prediction and Planning

This section introduces the G2P2C algorithm by presenting its neural network architecture, algorithmic improvements, and implementation details. This section uses the problem formulation (Section 4.2.1), clinical benchmarks (Section 4.2.3), and evaluation metrics (Section 4.2.4) presented in Chapter 4.

5.2.1 Architecture

G2P2C is a modular RL algorithm based on PPO, which introduces two additional optimisation phases, namely, **model learning** and **planning**. PPO was selected as the basis for designing G2P2C due to its demonstrated superior performance (Chapter 4) and efficiency in safety-critical applications, where excessive changes in the control policy could lead to unexpected behaviour. G2P2C is implemented using two separate neural networks with similar architecture; the Actor-Network (Π_θ) and the Critic-Network (V_ϕ). Each network has three modules as shown in Figure 5.2, a feature extractor module (E^Π, E^V), a glucose prediction module (M^Π, M^V) and a policy(π)/value(v) module respectively. The feature extractor modules E^Π and E^V are based on an LSTM network [Hochreiter and Schmidhuber, 1997], where the input is the observed state (s_n) and the output is the hidden state vector (h_n) of the LSTM. The Π_θ network included the policy module π, representing the policy func-

tion, while V_ϕ included the value module v, which represented the value function. They used h_n as the input. The output of the policy module (π) was formulated as a normal distribution $(\mathcal{N}(\mu, \sigma))$ over the action space, where both μ and σ parameters were learned. The output of the value module (v) was trained to predict the expected return $(v_\phi(s_t))$. The glucose prediction modules (M^Π and M^V) are trained to learn the glucose dynamics of the target T1D subject. The functionality of M^Π and M^V will be discussed in Section 5.2.2.2.

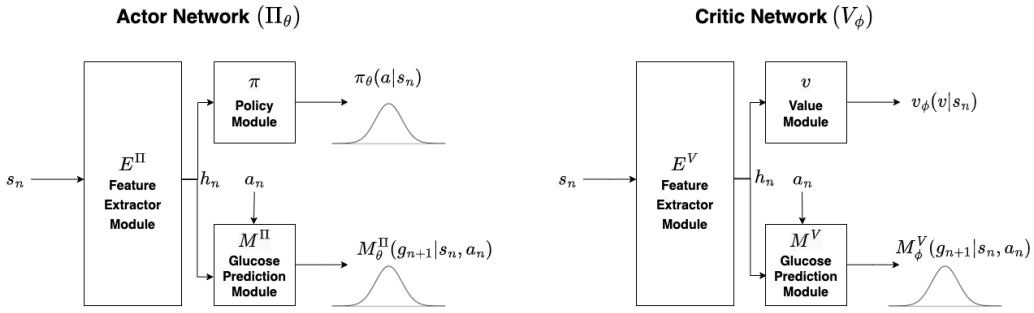


Figure 5.2: Schematic diagram of actor and critic networks.

5.2.2 Algorithm

G2P2C alternates between sampling and optimisation. During sampling, the current policy was used by w parallel agents and simulations were rolled out for n time steps. The resulting trajectory $(s_1, a_1, r_1, s_2, a_2 \cdots)$ information was stored in a data buffer (D). Once the sampling procedure was complete, the optimisation procedure commenced, consisting of three sequential update phases. The first phase used the standard policy and value update of PPO [Schulman et al., 2017], the second phase was the **model learning** update, and the third phase was the **planning** update. The second and third update phases were designed to extend the PPO algorithm to incorporate a short-horizon fine-tuning of the policy.

5.2.2.1 PPO Phase

During the first phase, the standard PPO optimisation update was carried out, where the optimisation objective of the policy module was to maximise the objective function $L^\pi(\theta)$ defined in Equation (5.1):

$$L^\pi(\theta) = \hat{\mathbb{E}}_t \left[min\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t, clip\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) + \beta_s H\Big(\pi(\cdot|s_t)\Big) \right], \tag{5.1}$$

where $\pi_{\theta_{old}}$ is the policy prior to the update. Excessive changes between the new and old policies were constrained by clipping the probability ratios of the policies to the interval $[1-\epsilon, 1+\epsilon]$. $H(\pi(\cdot|s_t))$ represents the entropy term used in the optimisation to facilitate exploration where β_s was a hyperparameter. The return G_t defined in Equation (4.7) was used to calculate the advantage function estimate $\hat{A}_t$ and value function targets $\hat{v}_t^{target}$. The value module objective was to minimise the objective function $L^v(\phi)$ defined in Equation (5.2).

$$L^v(\phi) = \hat{\mathbb{E}}_t \left[\frac{1}{2} \left(v_\phi(s_t) - \hat{v}_t^{target} \right)^2 \right]. \tag{5.2}$$

5.2.2.2 Model Learning Phase

The model learning phase succeeded the PPO phase. The glucose prediction modules M^Π and M^V introduced in the actor and critic networks, respectively, were used in this phase. As an auxiliary learning task, the M^Π and M^V modules were trained to learn the glucose dynamics of the target T1D subject. The hidden states of the LSTM (h_n) and a_n were integrated as inputs to the glucose prediction modules, where the output was the one-step ahead glucose estimate (g_{n+1}) represented as a normal distribution. This design was expected to further facilitate the learning of a dynamical system state representation $(s_n^\star)$ at the hidden state (h_n) of the LSTM networks, as hidden states (h_n) of LSTM networks are capable of learning a representation of the state space $(s_n^\star)$ of a dynamical system [Ljung et al., 2020]. The glucose prediction modules were implemented using the same architecture and trained for the same task to facilitate feature distillation between the networks, inspired by Cobbe et al. [2021]. Furthermore, the learned glucose dynamics model in the actor-network was later used in the planning phase.

A replay buffer (B) was introduced to store the latest trajectories experienced by the algorithm as triplets of s_t, a_t, and g_{t+1}. The model-learning update commenced once B was filled $(25,000$ interactions). The optimisation objective was based on maximum likelihood estimation, where the objective was to minimise $L^{M^\Pi}(\theta)$ and $L^{M^V}(\phi)$ defined in Equation (5.3), (5.4). The Kullback-Leiber divergence (d_{KL}) and Mean Squared Error (MSE) penalties applied in $L^{M^\Pi}(\theta)$ and $L^{M^V}(\phi)$ respectively aimed to minimise the divergence from the already learned policy $\pi_{\theta_{ppo}}$ and value function $v_{\phi_{ppo}}$ after the PPO update phase. Hyperparameters β_1 and β_2 were introduced to regularise the respective penalties.

$$L^{M^\Pi}(\theta) = \hat{\mathbb{E}}_t \left[-log\left(M_\theta^\Pi(g_{t+1}|s_t, a_t) \right) + \beta_1 d_{KL}\left[\pi_{\theta_{ppo}}(\cdot|s_t), \pi_\theta(\cdot|s_t) \right] \right]. \tag{5.3}$$

$$L^{M^V}(\phi) = \hat{\mathbb{E}}_t \left[-log\left(M_\phi^V(g_{t+1}|s_t, a_t) \right) + \beta_2 \frac{1}{2} \left(v_{\phi_{ppo}}(s_t) - \hat{v}_\phi(s_t) \right)^2 \right]. \tag{5.4}$$

5.2.2.3 Planning Phase

Following the model-learning phase, the third and final update was performed during the planning phase. A plan-space planning [Sutton and Barto, 2018] approach was introduced to improve the learned policy by integrating a short-term optimisation objective. The planning phase only used the Π_θ network since it focused on fine-tuning the learned policy module. Once M^Π achieved a prediction accuracy of Root Mean Squared Error (RMSE) $< e_{target}$ (15mg/dL), the planning phase commenced. M^Π was used to carry out m number of short-horizon ($n_{plan} = 6$ simulation steps (30 minutes)) Monte Carlo rollouts (τ) for each state stored in the buffer D. For each state, the rollout with the best *simulated-return* (τ^*) was identified according to Equation (5.5). The planning phase fine-tuned the policy toward the best action (a_t^*) associated with the target state (s_t). The planning objective was to minimise $L^{plan}(\theta)$, which was designed based on maximum likelihood estimation (Equation (5.6)). During the planning phase, the Π_θ network weights associated with E^Π and M^Π modules were kept fixed, and only the weights associated with the π module were updated. This was done to ensure that the planning phase reflected a fixed M^Π. The steps of the G2P2C algorithm are summarised in Algorithm 1.

$$\tau^* = \arg\max_{\tau} \left[(\sum_{q=1}^{n_{plan}} R_q) + v(s_{n_{plan}}) \right]. \tag{5.5}$$

$$L^{plan}(\theta) = \hat{\mathbb{E}}_t \left[-log\Big(\pi_\theta(a_t^*|s_t) \Big) \right]. \tag{5.6}$$

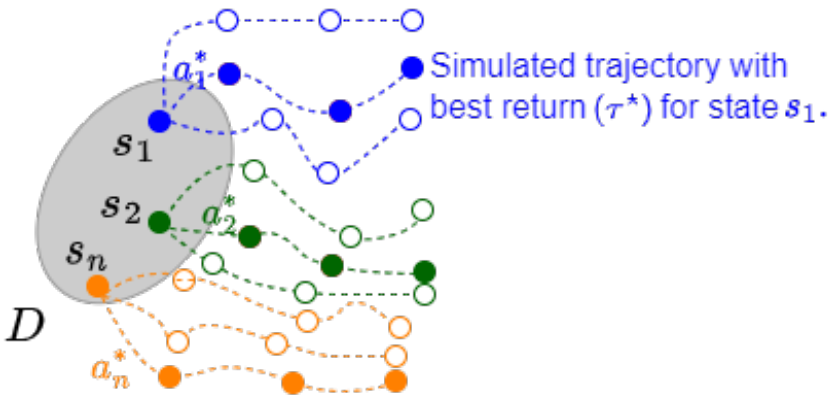


Figure 5.3: Illustration of the planning phase in G2P2C. During planning, short-horizon trajectories are simulated using the learned glucose dynamics model for the states stored in the rollout buffer D. Planning updates use the initial action with the best trajectory for each state.

Algorithm 1 G2P2C

Initialise an empty auxiliary buffer (B) of the size N_B.
Initialise the average reward estimate ($\hat{R}^{avg} = 0, N_{total} = 0$).
Initialise Actor-Network (Π) and Critic-Network (V) weights (θ, ϕ).
for $iteration = 1, 2, 3, \cdots$ **do**
  Initialise an empty data buffer (D) of the size N_D.
  Perform n-step rollouts for w parallel workers under current policy $\pi_{\theta_{old}}$ and store (s_n, a_n, r_n) transitions $\rightarrow D$.
  Store (s_n, a_n, g_{n+1}) transitions $\rightarrow B$.
  Compute the advantages $\hat{A}_t$ & value function targets $\hat{v}_t^{target}$.
  Update the average reward estimate ($\bar{R}_D = \frac{\sum r \in D}{N_D}$):
    $N_{total} \leftarrow N_{total} + N_D$,
    $\hat{R}^{avg} \leftarrow \hat{R}^{avg} + \frac{N_D}{N_{total}}(\bar{R}_D - \hat{R}^{avg})$.

  (1) PPO Phase:
  for $epoch = 1, 2, 3, \cdots, E_\pi$ **do**
    optimise L^Π wrt θ, on all data in D:
      $\theta \leftarrow \theta + \alpha_2 \cdot \nabla_\theta L^\Pi(\theta)$.
    Early stop:
      $d_{KL}[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)] > d_{target}$.
  end for
  for $epoch = 1, 2, 3, \cdots, E_V$ **do**
    optimise L^v wrt ϕ, on all data in D:
      $\phi \leftarrow \phi - \alpha_3 \cdot \nabla_\phi L^V(\theta)$.
  end for

  (2) Model Learning Phase:
  if B is filled **then**
    for $epoch = 1, 2, 3, \cdots, E_M$ **do**
      optimise L^{M^Π} wrt θ, on all data in B:
        $\theta \leftarrow \theta - \alpha_4 \cdot \nabla_\theta L^{M^\Pi}(\theta)$.
      optimise L^{M^V} wrt ϕ, on all data in B:
        $\phi \leftarrow \phi - \alpha_5 \cdot \nabla_\phi L^{M^V}(\phi)$.
    end for
  end if

  (3) Planning Phase:
  if $M^\Pi_{pred-error} < e_{target}$ **then**
    Fix parameters related to E^Π, M^Π modules of the actor network (Π).
    for $epoch = 1, 2, 3, \cdots, E_{plan}$ **do**
      optimise L^{plan} wrt θ, on all data in D:
        $\theta \leftarrow \theta - \alpha_5 \cdot \nabla_\theta L^{plan}(\theta)$.
    end for
  end if
  $\theta_{old} \leftarrow \theta$.
  $\phi_{old} \leftarrow \phi$.
  if $N_{total} > I_{total}$ **then**
    Stop.
  end if
end for

5.2.3 Implementation Details

To support comparisons, the neural network implementation and optimisation algorithm of G2P2C is similar to the state-of-the-art algorithm implementation presented in Section 4.2.2.4. The hyperparameters of G2P2C and their values are summarised in Table 5.1.

Table 5.1: Hyperparameters for G2P2C.

Hyperparameter	**Symbol**	**Value**
Sample time		5 minutes
Glucose sensor		Guardian RT
Insulin pump		Insulet
Augmented state history	k	12 (1-hour)
Total number of Interactions	I_{total}	$800,000$
Batch size of policy, value, model-learning, and planning		$1,024$
Number of steps per rollout	$n_{rollout}$	256
Number of workers	w	16
Data buffer (D) size	N_D	$n_{rollout} \cdot w$
Replay buffer (B) size	N_B	$25,000$
No. of policy epochs / value epochs / model-learning epochs	E_Π, E_V, E_M	5
No. of planning epochs	E_{plan}	1
Entropy Coefficient	β_s	0.001
Penalty Coefficient aux-policy / aux-value	β_1, β_2	0.01
Learning rate of policy / value / model-learning / planning	$\alpha_2, \alpha_3, \alpha_4, \alpha_5$	3 x 10^{-4}
PPO clip range	ϵ	0.1
Target Kullback-Leiber divergence (d_{KL}) threshold	d_{target}	0.01
Target glucose prediction error threshold	e_{target}	15mg/dL
Planning trajectories	m	50 (per state)
Planning horizon	n_{plan}	6 (30-minutes)

5.2.4 Experiments and Evaluation Metrics

The experimental procedure discussed in Section 4.2.5 was used to evaluate G2P2C. G2P2C used the same number of training interactions (800,000), which was assumed sufficient for the policies to converge. The similar experimental procedure and learning parameters provided fair comparisons between G2P2C and PPO.

The RL performance of G2P2C and PPO algorithms were analysed using their training curve characteristics and the Percentage Reward (PR) metric, as discussed in Sections 4.2.4 and 4.2.5. In addition, a modified version of the **post-convergence instability (PCI)** metric proposed in Dulac-Arnold et al. [2020] was used to compare the instability of the converged policies between PPO and G2P2C. To derive PCI, first, the evaluation simulations of both PPO and G2P2C (1,500 each) were used to calculate the 95% confidence interval ($[R^{eval}_{lower}, R^{eval}_{upper}]$) for R^{eval}, across the algorithms for each subject. Next, PCI was calculated as the percentage of evaluation simulations where R^{eval} is below R^{eval}_{lower}. Hence, $PCI(\pi) = \left(\frac{\sum_{i=1}^{1500} \mathbf{1}(R^{eval}_i < R^{eval}_{lower})}{1500}\right) * 100$, where $\mathbf{1}(.)$ was an indicator function. The objective was to minimise PCI, which reflects better stability. Furthermore, in addition to the cohort-level analysis, a statistical analysis was conducted to compare the PR performance between PPO and G2P2C algorithms for each *in-silico* subject accounting for the inter- and intra- population variability.

Following best practices, a cohort-level clinical analysis of G2P2C was conducted using the BBI and BBHE clinical treatment strategies and clinical metrics presented in Section 4.2.4, which included Time-In-Range (TIR) and Failure Rate (FR) metrics. A statistical analysis was conducted to compare the TIR performance of G2P2C against BBI and BBHE.

The statistical analysis were conducted using the IBM SPSS Statistics Software (Version-28.0.1.1). A Shapiro-Wilk Test [Shapiro and Wilk, 1965] was performed to check the normality, Mann-Whitney U Test [Mann and Whitney, 1947] was conducted to evaluate significance, and effect size calculated using the Pearson product-moment correlation coefficient (r) [Fritz et al., 2012].

5.3 Results

RL Analysis. G2P2C achieved a statistically significant ($P < 0.05$) PR improvement compared to PPO for 18 of the 20 subjects (Table 5.2). The performance improvement was non-significant for Adolescent0 and Adolescent4. Adolescent0 was the easiest to control under both RL-based methods (97.50%, 97.99%), leaving almost no margin for improvement for G2P2C. The inability of G2P2C to achieve significant PR improvements in Adolescent4 could be related to the observed instability compared to PPO, which should be further investigated. Adolescent6 (72.89%, 80.55%) was the hardest to control for both RL algorithms. G2P2C gave the largest improvement in the PR performance for Adolescent6, where it achieved a PR of 80.55% compared to 72.89% in PPO. Figure 5.4 summarises the PCI metric for the target subjects. PPO showed higher instability compared to G2P2C in all subjects except for Adult9 and Adolescent4. For some subjects (Adult0, Adult4, Adolescent1, Adolescent3, Adolescent6), G2P2C showed a very large improvement ($> 25\%$) in PCI.

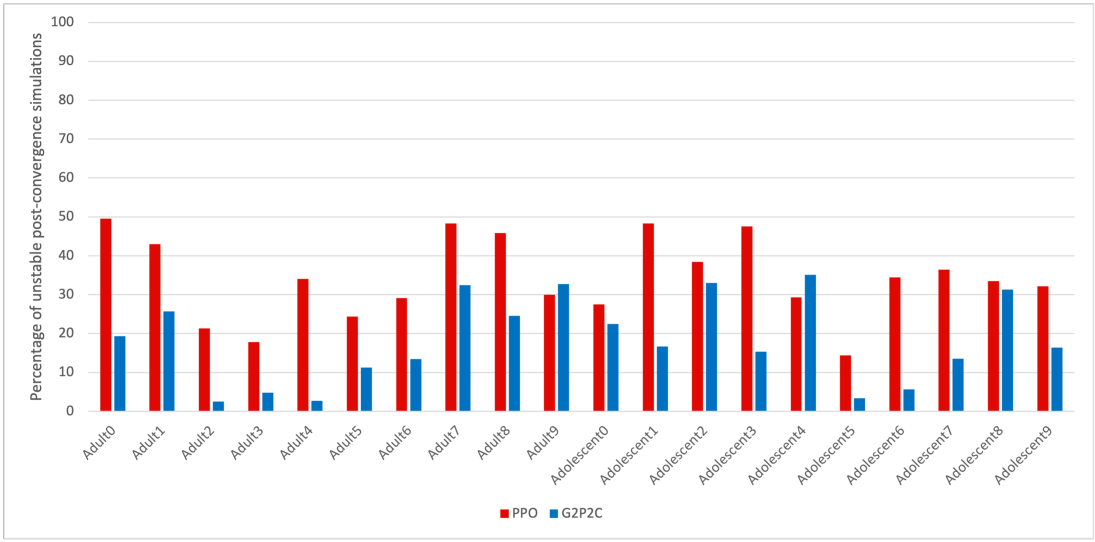


Figure 5.4: Post-convergence instability (PCI) comparison between PPO and G2P2C algorithms calculated based on evaluation simulations.

The standard deviation of the PR of the evaluation simulations was reduced in 18 out of the 20 subjects (except for Adult5 and Adult6) (Table 5.2); this behaviour is beneficial for the glucose control task to reduce uncertainty. The reduction in the standard deviation was also visible through a qualitative analysis of the reward curves. The reward curves present the learning behaviour of RL algorithms during training (Figures 5.5 and 5.6). The reduction in the standard deviation was attributable to the effect of the planning phase proposed in G2P2C. At the beginning of training, the reward curves had a similar standard deviation for both PPO and G2P2C (Figures 5.5 and 5.6). However, the planning phase of G2P2C was automatically initiated at approximately $200,000$ learning interactions once the learned glucose prediction module (M^{Π}) achieved the predefined accuracy threshold. By analysing the reward

curves, it can be observed that the standard deviation began to reduce in G2P2C compared to PPO once the planning was initiated. This reduction in the standard deviation was more prominent in some subjects (e.g., Adult0, Adolescent0) compared to others (e.g., Adult3, Adolescent6).

The experimental results indicate that changes in the learning horizon from long-term only (PPO) to additional short-term fine tuning (G2P2C) benefited toward the performance by reducing catastrophic failures. The fine-tuning of the policy to the short-term (Equation (5.6)) also reduced the variability in performance (Figure 5.5, 5.6). High variability in performance was observed among subjects (Adolescent0 was the easiest (Figure 5.8), while Adolescent6 was the hardest to control (Figure 5.8) (Table 5.2). Hence, an average reward RL objective was used in G2P2C, compared to the popular discounted RL setting, which requires personalised tuning of the discounting rate to learn satisfactory control policies.

Table 5.2: Comparison of percentage reward for all subjects based on evaluation simulations.

Subject	**PPO**	**G2P2C**	**Significance (PPO - G2P2C)**
Adult0*	94.00±4.42	95.50±1.02	$P < .001$, $r = 0.34$
Adult1*	95.69±3.39	96.31±0.99	$P < .001$, $r = 0.14$
Adult2*	87.33±14.822	91.92±6.53	$P < .001$, $r = 0.44$
Adult3*	88.71±12.13	90.64±9.86	$P < .001$, $r = 0.21$
Adult4*	92.40±9.21	94.73±5.03	$P < .001$, $r = 0.51$
Adult5*	88.17±9.49	88.19±14.80	$P < .001$, $r = 0.30$
Adult6*	92.07±7.51	92.32±9.91	$P < .001$, $r = 0.22$
Adult7*	94.77±3.39	95.57±1.77	$P < .001$, $r = 0.22$
Adult8*	89.82±6.26	91.18±1.41	$P < .001$, $r = 0.19$
Adult9*	92.46±5.85	92.85±1.08	$P < .001$, $r = 0.06$
Adolescent0	97.50±5.14	97.99±0.52	$P = .686$, $r = 0.01$
Adolescent1*	79.58±6.67	81.63±5.57	$P < .001$, $r = 0.30$
Adolescent2*	89.68±5.14	90.01±3.44	$P = .01$, $r = 0.05$
Adolescent3*	87.44±7.41	90.08±4.42	$P < .001$, $r = 0.42$
Adolescent4	87.98±7.48	88.17±3.77	$P = .831$, $r = 0.00$
Adolescent5*	90.09±11.12	91.10±9.34	$P < .001$, $r = 0.21$
Adolescent6*	72.89±12.57	80.55±7.59	$P < .001$, $r = 0.44$
Adolescent7*	80.52±9.11	82.67±3.15	$P < .001$, $r = 0.17$
Adolescent8*	92.99±4.33	93.10±2.06	$P < .001$, $r = 0.09$
Adolescent9*	86.60±9.42	87.81±8.55	$P < .001$, $r = 0.18$

* Statistical significance ($P < 0.05$), $N = 1,500$ simulations are conducted for each subject under each method. The effect size, $r > 0.1$: small effect, $0.3 < r < 0.5$: moderate effect, $r > 0.5$: large effect.
Acronyms: G2P2C: Glucose Control by Glucose Prediction and Planning, PPO: Proximal Policy Optimisation.

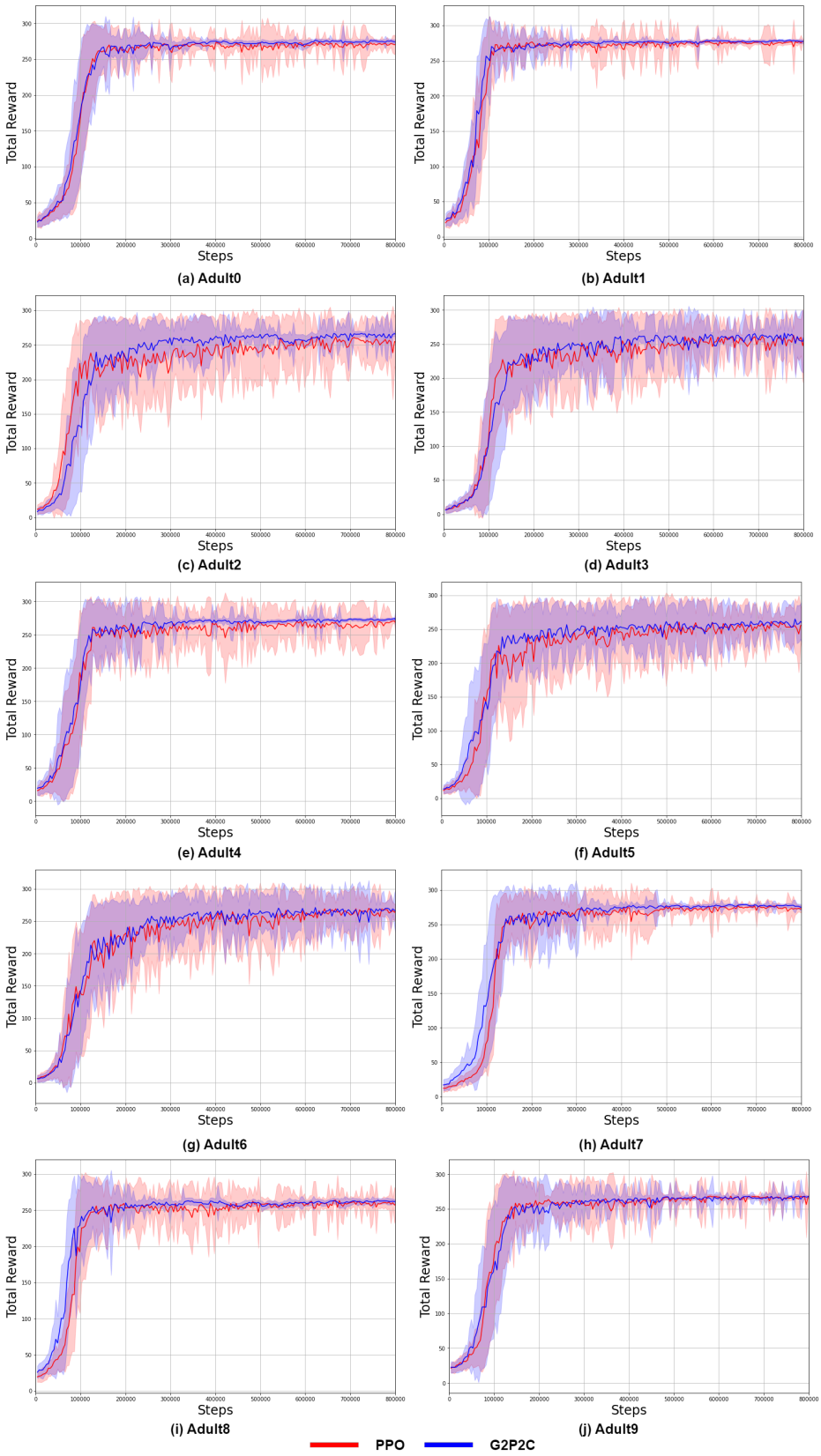


Figure 5.5: Comparison of PPO, G2P2C algorithms during training on the adult cohort. The mean and standard deviation (shaded area) of the total cumulative reward for evaluation simulations (3 random seeds x 20 evaluation scenarios / iteration) achieved is presented against 800,000 learning interactions.

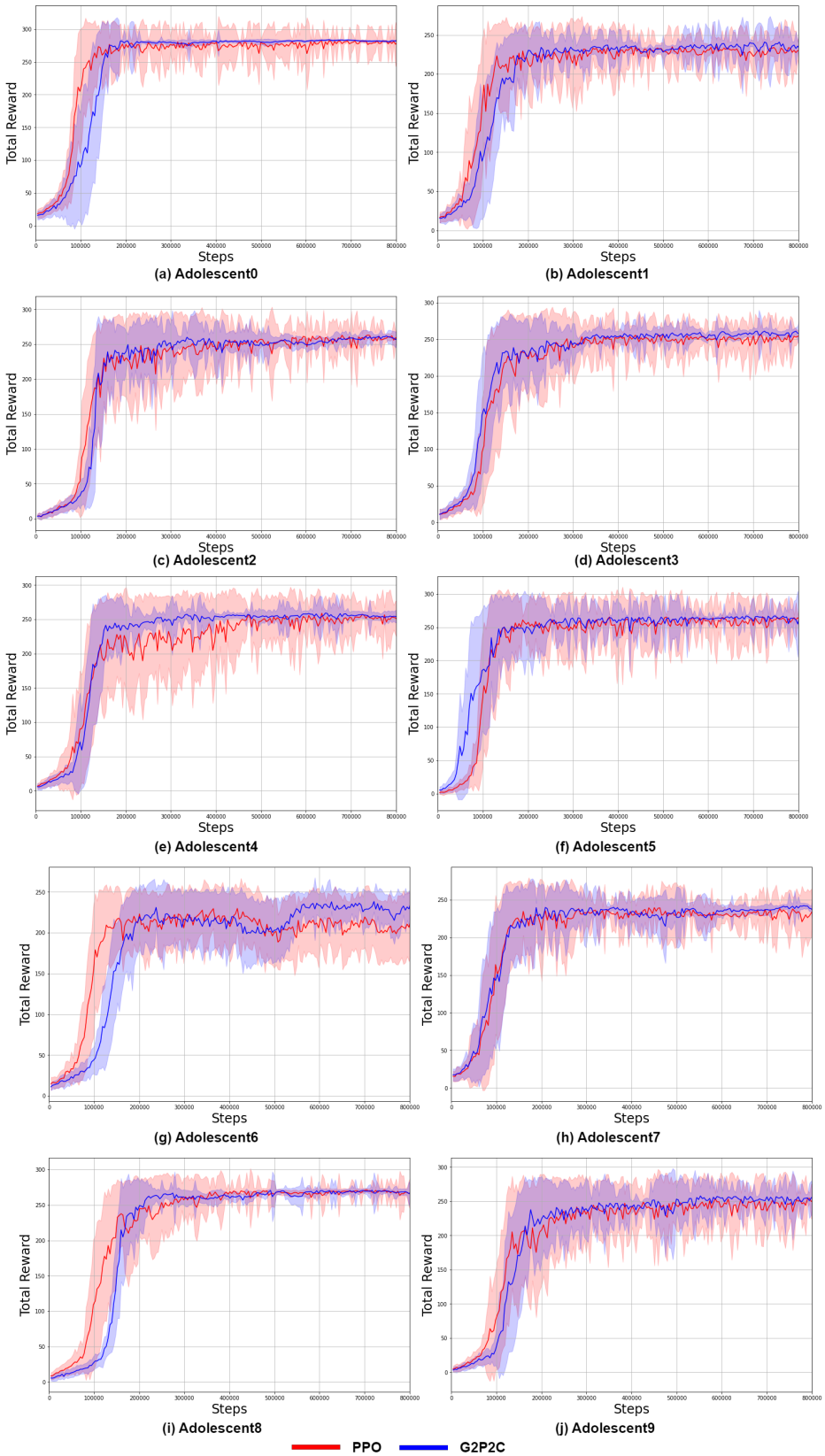


Figure 5.6: Comparison of PPO, G2P2C algorithms during training on the adolescent cohort. The mean and standard deviation (shaded area) of the total cumulative reward for evaluation simulations (3 random seeds x 20 evaluation scenarios / iteration) achieved is presented against 800,000 learning interactions.

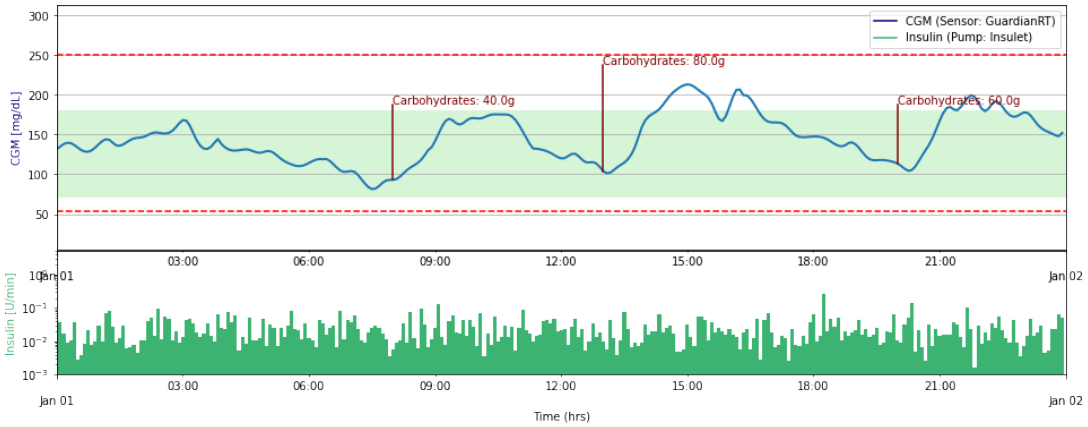


Figure 5.7: Sample glucose simulation trajectory for G2P2C based on Adult7.

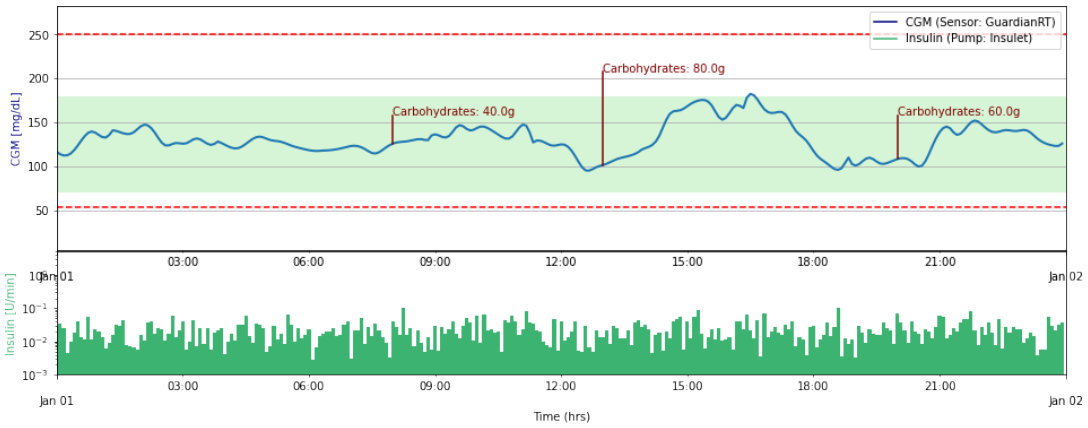


Figure 5.8: Sample glucose simulation trajectory for G2P2C based on Adolescent0.

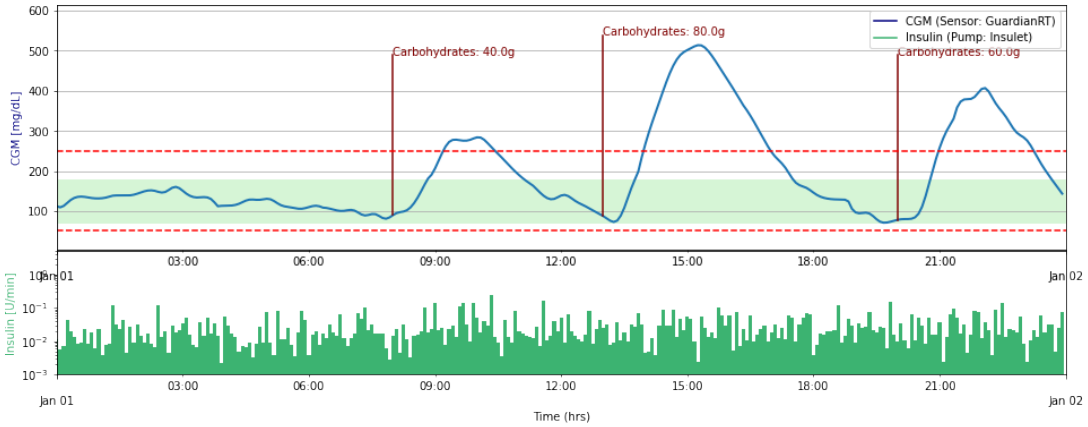


Figure 5.9: Sample glucose simulation trajectory for G2P2C based on Adolescent6.

Clinical Analysis. For the adult cohort, G2P2C achieved a mean TIR of 72.69%, which was higher than BBI (70.83%) and BBHE (69.79%). The improvement was statistically significant ($P < 0.05$) compared to BBI ($r = 0.09$) and BBHE ($r = 0.14$). The clinical performance of the candidate algorithms based on T1D-related criteria are summarised in Table 5.3, 5.4. The improvement of G2P2C compared to BB methods was mainly attributable to the reduction of the time in the hyperglycaemic range without an increase in the time in hypoglycaemic range. This is especially important considering the fact that G2P2C included no prior meal announcement. All algorithms (RL-based and BB treatment) presented comparable hyper- and hypoglycaemic risk profiles, as demonstrated by the HBGI and LBGI indices. In comparison to PPO, G2P2C achieved improved performance in all clinical metrics while also reducing the standard deviation. Finally, the RL-based algorithms showed a higher failure rate compared to the standard treatment methods. However, G2P2C managed to reduce the failure rate to 1.62% compared to 2.79% in PPO.

For the adolescent cohort, G2P2C achieved a mean TIR of 64.33% compared to 71.43% and 70.23% of BBI and BBHE. The difference was mainly observable in the time spent in the severe hyperglycaemia range, while the time in the hypoglycaemic zones was not largely affected. This was reflected in the hyper- and hypoglycaemic indices, with the RL-based algorithms maintaining a moderate-risk profile in terms of HGBI ($10.0 \leq \text{HGBI} \leq 15$ [Dalla Man et al., 2014]) and a low-risk profile in terms of LGBI($1.1 \leq \text{LGBI} \leq 2.5$ [Dalla Man et al., 2014]). The RL-based algorithms presented higher failure rates compared to the BB treatment methods. In this respect, the contribution of G2P2C markedly reduced failure rates compared to PPO with 1.48% and 4.93% failure rate respectively.

The observed TIR performance variability among *in-silico* subjects highlights the importance of conducting analysis comprising complete subject cohorts and multiple cohorts, which was identified as a limitation in previous work. Sample glucose simulation trajectories for G2P2C are presented in Figures 5.7, 5.8, and 5.9, while Appendix A provides sample trajectories for all 20 *in-silico* subjects (an online demonstration is introduced in Chapter 8, which can be used to conduct custom simulations to generate sample glucose trajectories).

Table 5.3: Clinical performance comparison of G2P2C against the state-of-the-art RL and basal-bolus treatment for the adult cohort.

Method	Failure (%)	Severe Hypo.(%) TBR-Level 2	Hypo. (%) TBR-Level 1	Normo. (%) TIR	Hyper. (%) TBR-Level 1	Severe Hyper.(%) TBR-Level 2	RI	LBGI	HBGI
BBI	0.39	0.00†	0.00	70.83	22.57	2.08	7.68	0.81	6.72
		0.00-0.00*	0.00-0.00	61.46-79.17	17.36-31.94	0.00-4.86	6.07-9.38	0.28-1.68	4.97-7.94
		0.06(0.40)‡	0.64(2.01)	71.02(11.29)	24.44(10.57)	3.85(5.72)	8.35(4.05)	1.33(1.60)	7.02(2.79)
BBHE	0.35	0.00	0.00	69.79	23.61	3.12	7.62	0.41	6.92
		0.00-0.00	0.00-0.00	60.42-78.47	18.06-32.29	0.00-5.21	5.78-8.69	0.11-0.93	5.11-8.11
		0.05(0.36)	0.40(1.42)	69.78(11.29)	25.44(10.57)	4.33(5.64)	8.00(3.74)	0.88(1.37)	7.11(2.67)
PPO	2.79	0.00	0.00	69.44	20.83	7.99	9.56	0.89	8.05
		0.00-0.00	0.00-1.04	62.15-76.04	16.32-25.35	2.43-12.85	6.89-12.04	0.28-2.17	5.86-10.10
		0.32(1.51)	1.19(2.77)	69.12(10.53)	20.93(7.11)	8.44(6.88)	9.79(3.66)	1.64(2.11)	8.14(3.05)
G2P2C	1.62	0.00	0.00	72.57	18.75	6.60	9.00	1.04	7.42
		0.00-0.00	0.00-1.04	66.32-79.86	15.28-23.26	0.69-10.76	6.21-11.04	0.43-2.14	5.18-9.35
		0.22(1.26)	1.11(2.50)	72.69(9.53)	19.37(6.46)	6.61(5.53)	8.94(3.18)	1.58(1.74)	7.36(2.60)

The †median, *inter-quartile range followed by the ‡mean (standard deviation) are presented. Acronyms: BBHE - Basal Bolus Human Error, BBI - Basal Bolus Ideal, G2P2C - Glucose Control by Glucose Prediction & Planning, HBGI - High Blood Glucose Index, LBGI - Low Blood Glucose Index, PPO - Proximal Policy Optimisation, RI - Risk Index, RL - Reinforcement Learning, TAR - Time Above Range, TBR - Time Below Range, TIR - Time In Range.

Table 5.4: Clinical performance comparison of G2P2C against the state-of-the-art RL and basal-bolus treatment for the adolescent cohort.

Method	Failure (%)	Severe Hypo.(%) TBR-Level 2	Hypo. (%) TBR-Level 1	Normo. (%) TIR	Hyper. (%) TBR-Level 1	Severe Hyper.(%) TBR-Level 2	RI	LBGI	HBGI
BBI	0.00	0.00†	0.00	67.71	16.32	4.86	7.93	0.43	7.56
		0.00-0.00*	0.00-0.00	63.54-76.39	11.11-27.43	0.00-18.40	5.21-14.10	0.09-1.40	5.04-12.26
		0.03(0.27)‡	0.44(1.27)	71.43(12.31)	19.65(11.53)	8.47(8.82)	9.26(4.83)	1.07(1.57)	8.19(3.78)
BBHE	0.00	0.00	0.00	66.67	18.06	5.56	8.06	0.21	7.94
		0.00-0.00	0.00-0.00	63.19-73.26	11.11-28.82	1.74-18.75	5.68-14.11	0.01-0.87	5.67-12.59
		0.01(0.10)	0.20(0.76)	70.23(12.52)	20.37(12.04)	9.20(8.85)	9.15(4.51)	0.69(1.11)	8.46(3.82)
PPO	4.93	0.00	0.00	60.42	15.62	17.71	14.66	1.21	12.75
		0.00-0.00	0.00-1.39	54.17-70.14	12.15-20.83	11.11-27.43	11.16-20.63	0.49-2.45	9.73-18.84
		0.28(1.36)	1.30(2.78)	63.72(13.95)	16.15(7.23)	18.55(11.09)	15.40(6.67)	1.82(1.99)	13.58(6.55)
G2P2C	1.48	0.00	0.00	60.76	15.97	15.97	14.29	1.17	12.24
		0.00-0.00	0.00-1.04	55.56-70.14	12.50-21.88	10.07-28.12	11.20-20.74	0.52-2.26	9.75-19.48
		0.17(0.95)	1.07(2.35)	64.33(13.18)	16.74(7.17)	17.69(10.51)	15.10(6.50)	1.65(1.64)	13.45(6.23)

The †median, *inter-quartile range followed by the ‡mean (standard deviation) are presented. Acronyms: BBHE - Basal Bolus Human Error, BBI - Basal Bolus Ideal, G2P2C - Glucose Control by Glucose Prediction & Planning, HBGI - High Blood Glucose Index, LBGI - Low Blood Glucose Index, PPO - Proximal Policy Optimisation, RI - Risk Index, RL - Reinforcement Learning, TAR - Time Above Range, TBR - Time Below Range, TIR - Time In Range.

5.4 Discussion

Characteristics of G2P2C. In this chapter, a novel RL-based algorithm, named G2P2C, for glucose control in T1D was proposed. G2P2C is capable of (1) providing fully-automated insulin infusion estimates, including both basal and bolus, and (2) does not require meal announcement and CHO estimation. The PPO algorithm was selected as the base algorithm to develop G2P2C due to its demonstrated efficient and stable learning achieved through a clipped optimisation objective which is beneficial for safety-critical applications [Schulman et al., 2017]. However, in glucose control experiments using the PPO algorithm, large unwanted glucose fluctuations were observed in the short term, ultimately leading to catastrophic failures. Hence, two novel optimisation phases, model learning and planning, were introduced in G2P2C. The model learning phase aimed to learn a glucose dynamics model, while the planning phase utilised the learned model to simulate short-term trajectories and fine-tuned the learnt policy for the short horizon.

G2P2C improved the sample efficiency by re-using collected experiences and facilitating a simulation-based exploration of the glucose state space using the learned model. This should be beneficial in real-world online learning applications where real experience is limited to tuning the policy. For example, during the transfer of the *in-silico* learned policy to *in-vivo*, a learned glucose dynamics model of the person may be used to simulate experience for the fine-tuning of the RL algorithm. The learned model is also expected to be valuable in feature distillation between the actor and value networks and in designing safety modules for an APS where the glucose dynamics model can be used to predict potential high and low blood glucose events targeting safety. The control performance and algorithmic characteristics of G2P2C show promise as a candidate algorithm for glucose control in APS.

Comparison with State-Of-The-Art RL Algorithms. In Chapter 4, the performance of PPO was identified to be superior compared to the other state-of-the-art RL algorithms. Hence, this chapter compared the performance of G2P2C against PPO. Through the proposed mechanism, G2P2C improved safety by reducing the catastrophic failures to 1.62% and 1.48% in the adult and adolescent cohorts, respectively, compared to 2.79% and 4.93% in PPO. G2P2C also improved the TIR performance compared to PPO in both cohorts by achieving 72.69% and 64.33% compared to 69.12% and 63.72% in the adult and adolescent cohorts respectively.

Comparison with Clinical Treatment Methods. The clinical performance of G2P2C was evaluated based on several T1D-related metrics and compared against standard treatment methods commonly used in clinical practice. G2P2C achieved a higher TIR of 72.69% for the adult cohort compared to BBI (71.02%) and BBHE (69.78%) while eliminating the need for CHO estimation and meal announcement, as well as corrective boli, upon which both BB methods rely. For the case of the adolescent cohort, which is much more challenging than the adult cohort, the TIR achieved by G2P2C was lower than the BB methods without and with human error (64% compared to

71.43% and 70.23% respectively). However, the performance comparison should account for G2P2C receiving no prior information related to upcoming meals.

Comparison with Previous Work. Research on glucose control often benchmarks the performance of algorithms against standard clinical treatment approaches and guidelines, as presented above in this work. Comparisons with prior work are rare and complicated due to the different cohorts, protocols, and simulator versions used. Despite differences in their methodology and for the sake of a more thorough assessment, a comparison of the proposed algorithm's performance with previous work in terms of TIR is presented in Table 5.5. The comparison includes RL-based fully autonomous systems proposed by Fox and Wiens [2019]; Fox et al. [2020]; Lee et al. [2020b] and hybrid systems by Lim et al. [2021]; Hettiarachchi et al. [2022b].

In order to make meaningful comparisons with previous work, it is important to first discuss the essential differences among the presented studies. One of the main differences relates to the the type of population used for the algorithmic assessment. Some previous studies have only considered handpicked individual subjects [Fox and Wiens, 2019] or adult-only cohorts [Lee et al., 2020b]. Fox et al. [2020] has considered child, adult and adolescent cohorts, however, they report the overall median TIR values. Lim et al. [2021] have used the adolescent and adult cohorts and present the mean TIR for each cohort. This thesis have also used both adolescent and adult cohorts and assessed the algorithmic performance separately for each. This is particularly important as it was observed that the adolescent cohort is much harder to control than the adult cohort. Another important difference among studies is related to the level of complexity of the testing meal protocol. Some studies have used meal protocols with **reasonably low CHO contents**[1] [Fox and Wiens, 2019; Fox et al., 2020] while others have focused on more challenging meal protocols with 180g of CHO daily [Lee et al., 2020b; Lim et al., 2021]. Similar to those, in this work, a high CHO meal scenario was used.

All the reported studies have used the UVA/PADOVA simulator, which was originally designed based on the MATLAB framework [Li, 2001]. However, previous studies have used different versions and implementations of the simulator. This is mainly due to the lack of support extended by existing simulators for the integration with the Python framework [vanRossum and Drake, 2010] which is predominantly used for designing RL algorithms due to its favourable characteristics (e.g., the ability to simulate parallel environments and use existing machine learning frameworks). Hence, Lim et al. [2021] and Lee et al. [2020b] have used independent customised platforms based on the UVA/PADOVA simulator while Fox and Wiens [2019]; Fox et al. [2020] has adopted an open-source version. In this thesis, the open-source version was used to ensure the reproducibility.

[1] These studies do not explicitly present the meal protocol used; instead, they emphasise that the CHO content was reasonably low. By further investigating the provided pseudocode, it was identified that the meals used were personalised based on the subjects' body weight. The protocol also included snacks in addition to the three main meals. However, the average daily CHO content was reasonably lower than 180g. A protocol that personalises meals complicates the evaluation of the system as low-

Table 5.5: Comparison of G2P2C against previous work on RL-based APS.

Approach	Adults (TIR)	Adolescents (TIR)	Average daily CHO (g)	Simulator
Fox and Wiens [2019]				
SAC†	Full cohort not considered		Low CHO Meals (values not provided)	Simglucose 2018 (UVA/PADOVA 2008)
Fox et al. [2020][a]				
SAC (RL-MA)*	77.12		Low CHO Meals (values not provided)	Simglucose 2018 (UVA/PADOVA 2008)
SAC (RL-Scratch)†	72.68			
Lee et al. [2020b]				
PPO†	89.30±4.19	-	180	Custom platform based on UVA/PADOVA 2013
Lim et al. [2021][b]				
SAC*	65.93±17.29	62.20±19.99	180 (157.5-202.5)	Custom platform based on UVA/PADOVA 2013
Hettiarachchi et al. [2022b][c]				
PPO*	65.02	-	294	Simglucose 2018 (UVA/PADOVA 2008)
This thesis				
PPO†	69.12±10.53	63.73±13.95	180	Simglucose 2018 (UVA/PADOVA 2008)
G2P2C†	72.69±9.53	64.33±13.18		

The TIR results are presented as mean±std for Lee et al. [2020b], Lim et al. [2021], and this thesis. The median TIR is presented for Fox et al. [2020]. [a]Results presented as the median TIR for all cohorts. This is the only RL-based study with the Child cohort. [b]Meal intake information required, scenario with multiple small meals is used ($40g$, $50g$, $20g$, $50g$, $20g$ ±12.5%).[c]A study conducted during this thesis where a hybrid system was designed and evaluated based on a very large CHO meal protocol (3 snacks were included in addition to the main meals).
*Hybrid system (require at least meal announcement). † Fully autonomous system. Acronyms: CHO - Carbohydrates, G2P2C - Glucose Control by Glucose Prediction & Planning, PPO - Proximal Policy Optimisation, TIR - Time In Range.

Lim et al. [2021] and Lee et al. [2020b] are the most comparable to this thesis due to the similar challenging meal protocol used[2]. Compared to Lim et al. [2021], which is a hybrid approach requiring meal announcement and CHO estimation, G2P2C improved the performance in both the adult and adolescent cohorts. However, the performance on the adult cohort was less than Lee et al. [2020b]. The evaluation trials conducted were different in these studies, where Lim et al. [2021] used a 10-day simulation for each subject without multiple repetitions, while Lee et al. [2020b] conducted seven random daily trials, as opposed to 1,500 random daily trials performed in this study. The *catastrophic failure* rate was not presented in Lim et al. [2021] and Lee et al. [2020b] studies, while Fox et al. [2020] reported a FR of 0%. However, they defined failures as simulations where glucose levels are $\leq$ 5 mg/dL. In contrast, in our work we defined a much stricter FR where glucose levels $\leq$ 40 mg/dL or $\geq$ 600 mg/dL were considered failures[3]. The lack of appropriate benchmarks in this field of research and the unavailability of source code and hyperparameters of previous work are hurdles towards meaningful comparisons and are discussed further in Section 9.2. To facilitate future research in this domain and for reproducibility, this work's implementations and hyperparameter settings are published on GitHub (https://github.com/chirathyh/G2P2C) as open-source under the MIT license.

CHO could be used to offset the performance reduction for the more challenging subjects of the cohort.

[2]Previous work on APS have mainly focused on meal scenarios with 3 main meals with challenging CHO contents as used in this thesis. However, RL-based APS studies have also explored scenarios with multiple small meals. A sub-study conducted during this thesis Table 5.5 [Hettiarachchi et al., 2022b] also explored the use of additional snacks. Glucose control is a uni-directional control problem as insulin is used only to reduce glucose (Section 2.4). The meals have an opposite effect on glucose. Hence, having multiple small meals spread throughout the day may help the RL algorithm in achieving better performance, as the opportunity for hypoglycaemia is less, making it difficult to assess the true performance. Therefore the main body of work presented in this thesis follows the widely adopted meal scenario with 3 main meals.

[3]A glucose level $\leq$ 40 mg/dL is life-threatening and can result in major cardiovascular and cerebrovascular problems [Kalra et al., 2013]. A glucose value $\geq$ 600 mg/dL is also life-threatening emergency and referred to as the Hyperosmolar hyperglycaemic state [Stoner, 2005]. The detectable range of current glucose sensors is roughly 40–600 mg/dL. [Kovatchev et al., 2009]

Chapter 6

Non-linear Continuous Action Spaces for RL in Glucose Control

"One of the challenges that arise in RL, and not in other kinds of learning, is the trade-off between exploration and exploitation...The agent has to exploit what it has already experienced in order to obtain reward, but it also has to explore in order to make better action selections in the future."
[Sutton and Barto, 2018, p.3].

Continuous action spaces are a challenge faced by RL algorithms when applied to real-world problems (Chapter 3). In the case of glucose regulation, a large continuous action space is required to reflect the dynamic insulin needs. The standard approach in RL is to formulate a linear action space for the agent. Therefore, this chapter introduces non-linear continuous action spaces to overcome the challenge associated with efficiently exploring the dynamic range of insulin in glucose control. Three non-linear translation functions are proposed to map the RL action to the insulin infusion rate, inspired by the basal-bolus pattern of clinical insulin treatment practice. The proposed non-linear action spaces are compared and evaluated against the standard linear action space used in practice for RL algorithms. An action space designed through this proposed novel approach is used in the RL-based APS presented in Chapters 4 and 5.

6.1 Introduction

This chapter introduces a study conducted to design the action space presented in the problem formulation of Chapter 4. The action space design proposed in this chapter is used in all RL-based APS presented in this thesis (Chapters 4 and 5). The action space is a vital element as it reflects the set of all possible actions available for the RL agent. The actions available in APS are insulin infusion rates. One of the main challenges RL algorithms face in the APS context is the large and continuous insulin action space, which differs from the discrete actions present in board game environments.

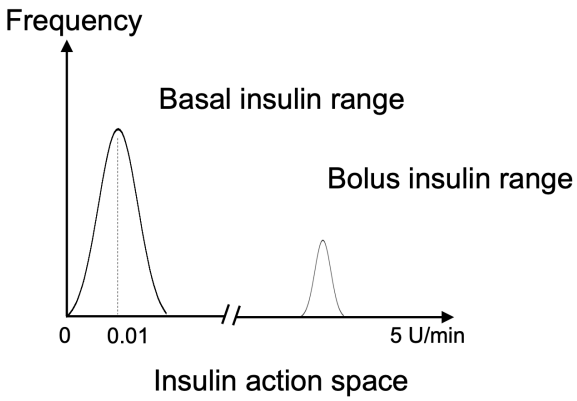


Figure 6.1: Frequency distribution of insulin action based on a clinical perspective (not to scale).

The basal-bolus strategy is a standard clinical treatment for T1D [Diabetes Control and Complications Trial Research Group, 1993]. According to the basal-bolus scheme, low-range basal insulin actions account for the vast majority of the total insulin actions. In contrast, the large bolus actions are intermittent. Hence, typical clinical patterns of insulin delivery can be interpreted as bi-modal in frequency (Figure 6.1). This challenges the RL algorithm since it is required to efficiently explore the entire continuous insulin action space to learn suitable control strategies for different situations requiring varying amounts of insulin out of a large dynamic range.

Applying RL algorithms to continuous action spaces is not straightforward compared to low-dimensional discrete action spaces, while high-dimensional actions further increase the learning difficulty [Lillicrap et al., 2015]. The complexity of continuous action spaces can be sub-optimally solved by discretising the action space. However, this may not be suitable for high-precision control problems such as glucose regulation, as it could eliminate the required information regarding the structure of the action space. Current available CSII pumps are discretised with fine resolution (e.g., Medtronic Minimed Pump basal increment of 0.025U/hr [Insulin Pump Comparison, 2022 [Online]]), allowing the assumption of a continuous insulin action space. An alternative approach could be the introduction of two separate actions for the RL algorithm to focus on the clinical conventions of basal and bolus insulin separately. However, this increases the degrees of freedom in the algorithm and could add complexity to learning due to the very sparse use of large insulin doses.

According to the RL-related literature, applications with a large continuous actions space can present challenges related to its efficient exploration [Lazaric et al., 2007] and the existence of redundant or irrelevant actions [Zahavy et al., 2018]. Lazaric et al. [2007] used a novel actor-critic algorithm based on a Sequential Monte Carlo approach to improving the exploration, while Zahavy et al. [2018] proposed an approach that combined the RL algorithm with an action elimination network to eliminate sub-optimal actions.

In the problem of glucose regulation in T1D, the majority of previous RL-based

approaches focus on hybrid systems, which only control basal insulin levels, while bolus insulin infusion is carried out manually by the user [Zhu et al., 2019, 2020; Fox and Wiens, 2019; Myhre et al., 2018; Ngo et al., 2018a,b; Lim et al., 2021]. These studies used a small set of handcrafted discrete actions to represent basal insulin [Zhu et al., 2019, 2020]. Only a handful of studies have sought to control both basal and bolus insulin without any user input [Fox and Wiens, 2019; Fox et al., 2020; Lee et al., 2020b]. In particular, Lee et al. [2020b] and Fox and Wiens [2019] used a discrete action distribution to control insulin. However, a continuous action space is expected to enable more flexible RL agents that can learn more robust control strategies [Fox and Wiens, 2019]. Fox et al. [2020] is the only research that focused on a continuous insulin action space. They used a tanh function in the final calculation of the RL action and encouraged a sparse insulin treatment strategy by translating negative values of the tanh function as no insulin administration, while the positive values were mapped linearly to the insulin pump action. The system was evaluated on reasonably low CHO meals, which required insulin rates $<$0.5 U/min. However, in real-life scenarios, the presence of meals with large CHO content results in a large insulin action space ([0, 5] U/min).

This thesis focuses on developing fully automated RL-based APS capable of handling challenging meal protocols, which include meals of large CHO contents. This translates to the need for a large insulin action space reflective of a real-life scenario. To address this challenge, non-linear action space representations are proposed for the RL algorithm. In contrast, the common practice in RL-based continuous control tasks is to use linear action spaces [Brockman et al., 2016; Tassa et al., 2018; Todorov et al., 2012]. The proposed non-linear action spaces provide a non-uniform resolution across the action space, which guides the RL algorithm to explore insulin delivery patterns observed in clinical treatment. Therefore, this chapter evaluates the impact of using non-linear action spaces in an RL problem formulation to improve performance, efficiency and quality of learning.

6.2 Methodology: Action Space Representations

The RL problem is formulated as a POMDP according to the state space and reward function designs presented in Chapter 4. In this work, different action space representations are evaluated to address the above-identified challenges while ensuring that the RL algorithm and hyperparameter settings are kept fixed. The PPO algorithm is selected as the RL agent. The task of the RL agent is to predict an action ($a \in [-1, 1]$) for a given state, which is mapped to the insulin infusion rate of the insulin pump ($I_{pump} \in [0, 5]$ U/min) based on a translation function T. Hence, T represents a mapping between the action space of the RL agent to the control space of the insulin pump (Equation (6.1)).

$$I_{pump} = T(a), a \in [-1, 1], I_{pump} \in [0, I_{max}]. \tag{6.1}$$

Glucose regulation requires frequent use of very small insulin doses for basal

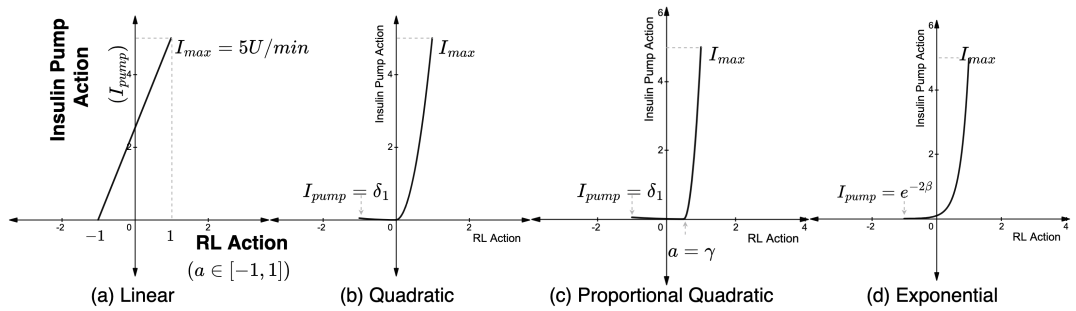


Figure 6.2: Translation functions used to map RL action to the insulin infusion rate: (a) linear, (b) quadratic, (c) proportional-quadratic, and (d) exponential.

insulin compared to the less frequent larger doses, resulting in a skewed concentration in the action space, as opposed to the uniform resolution provided by the linear mapping (Equation (6.2)). In order to capture this property, we explore three non-linear translation functions; quadratic, proportional-quadratic, and exponential to formulate non-linear action spaces in order to provide better resolution to the important target insulin ranges and use a standard linear action space for comparisons (Figure 6.2).

6.2.1 Linear Translation

The linear mapping of RL action to the control space is described in Equation (6.2), where I_{max} corresponds to the maximum insulin. This formulation is consistent with standard RL practice [Brockman et al., 2016; Tassa et al., 2018; Todorov et al., 2012].

$$I_{pump} = I_{max} \cdot \frac{(a+1)}{2}, a \in [-1, 1]. \tag{6.2}$$

6.2.2 Quadratic Translation

The translation function T is a quadratic function of the RL action a. This formulation integrates the two distinct actions (basal-bolus) used in typical insulin treatment to a single continuous action space avoiding the complexity of using multiple actions. The action space is divided into two segments, where actions in [-1, 0] are translated to a basal dose range with a maximum basal insulin of $\delta_1(0.05)$ and actions in (0,1] considered as the bolus range with a maximum bolus insulin of I_{max}. This results in a duplication of the basal range [0, 0.05] in the bolus range [0, 5], which could be considered negligible due to the low resolution of the bolus range.

$$I_{pump} = \begin{cases} \delta_1 \cdot a^2 & -1 \leq a \leq 0 \\ I_{max} \cdot a^2 & 0 < a \leq 1 \end{cases}. \tag{6.3}$$

6.2.3 Proportional-Quadratic Translation

This function is a modification of the Quadratic function where the parameter $\gamma(0.5)$ is introduced to adjust the resolution of the two dose ranges and $\delta_1(0.5)$ is the maximum basal insulin.

$$I_{pump} = \begin{cases} \frac{\delta_1}{(\gamma+1)^2} \cdot (a-\gamma)^2 & -1 \leq a \leq \gamma \\ \frac{I_{max}}{(\gamma-1)^2} \cdot (a-\gamma)^2 & \gamma < a \leq 1 \end{cases}. \tag{6.4}$$

6.2.4 Exponential Translation

The translation function T is an exponential function of the RL action $a \in [-1, 1]$ with a tuneable parameter $\beta(4.0)$ which ensures $I_{pump} \in (0, 5]$. This increases the resolution of the basal dose range while ensuring the action space is continuous without any duplication of actions. This formulation provides more flexibility for the RL algorithm to use the fully continuous structure of the action space and avoids any instabilities in learning, which might be caused by action duplication.

$$I_{pump} = I_{max} \cdot e^{\beta(a-1)}, a \in [-1, 1]. \tag{6.5}$$

6.2.5 Experiments and Evaluation Metrics

The simulator setup, training protocol, and evaluation protocol used in Chapter 4 were used to conduct the experiments. The challenging adolescent T1D cohort of the simulator and the PPO algorithm was selected as the RL agent for the analysis. Each action space representation was evaluated for three random seeds per subject, where all other hyperparameters were kept fixed. The PPO algorithms were trained for $500,000$ interactions ($1,736$ human days, 1 interaction = insulin action taken every 5 minutes), which was identified as sufficient to reach convergence. Upon the conclusion of training, $1,500$ evaluation simulations were also conducted for each subject.

The Percentage Reward (PR), Time In Range (TIR) and Failure Rate (FR) metrics introduced in Chapter 4 were used to compare the candidate action spaces based on their evaluation simulations. The best-performing action space representation in terms of PR was compared against the benchmark linear action space by conducting statistical significance tests for each individual subject. A Shapiro-Wilk Test [Shapiro and Wilk, 1965] was performed to check the normality, and a Mann-Whitney U Test

[Mann and Whitney, 1947] was conducted to evaluate significance using a confidence level of 0.05.

The training simulations were used to analyse the learning efficiency and quality of the candidate action spaces. The aim of the RL algorithm during training is to iteratively improve a control policy reasonably fast (relative to the application's time scale) while avoiding excessive changes (smooth learning), which could lead to sudden unanticipated behaviour and even result in catastrophic failures. To evaluate the learning efficiency, the **reward thresholds** of 25%, 50%, 70%, and 80% of the maximum achievable reward ($R^{\star}$) were defined. The average number of interactions required to reach the threshold as a **percentage of total interactions (PI)** was used to compare the learning efficiency between the candidate action spaces. Furthermore, through visual inspection of the shape and fluctuations of the reward curves during training, the learning smoothness of the RL agents was qualitatively assessed.

6.3 Results

All the proposed non-linear action spaces outperformed the linear action space. The exponential action space improved the performance in terms of PR by 24%, while reducing the FR by 42% compared to the benchmark linear action space, on average, across the subjects. The improvement was statistically significant ($p < .001$) for all subjects. The performance of the candidate action spaces for all subjects, based on PR and FR metrics, is summarised in Table 6.1. An inter-subject variability in performance improvements was identified, with substantial performance improvements for some subjects (Adolescent8). The proposed non-linear functions all performed similarly in terms of TIR and FR (Figure 6.3). Figure 6.3 also highlighted the variability in glucose control performance present among the subjects.

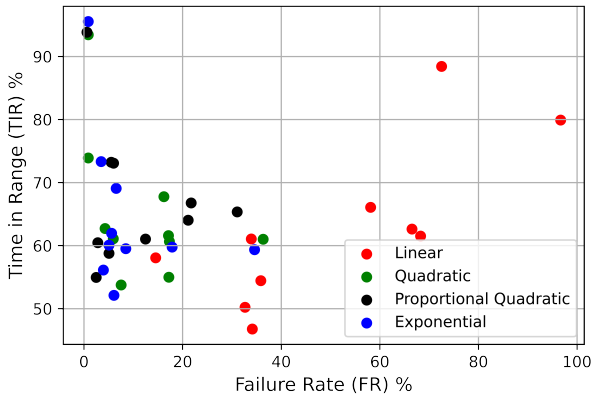


Figure 6.3: Time-In-Range (TIR) and Failure Rate (FR) for candidate action spaces (Each dot represents an adolescent subject and each color a candidate action space).

The exponential action space was the most efficient in reaching the 80% reward threshold in 34.90% PI, for all 10 subjects. Table 6.2 summarises the PI required for

Table 6.1: Adolescent cohort summary results — Percentage Reward (PR) & Failure Rate (FR) for the candidate action spaces: Linear(L), Quadratic(Q), Proportional-Quadratic(PQ), Exponential(E).

Adolescent ID	**Reward (*PR*)**				**Failure Rate (*FR*)**			
	L	***Q***	***PQ***	***E***	***L***	***Q***	***PQ***	***E***
0	52.55%	97.08%	97.33%	97.52%	72.53%	0.87%	0.60%	0.87%
1	66.70%	79.23%	81.75%	79.36%	35.87%	7.53%	2.47%	6.07%
2	76.09%	80.82%	77.92%	87.71%	32.67%	17.33%	21.73%	5.60%
3	58.48%	88.27%	75.20%	86.95%	34.13%	4.27%	31.07%	5.07%
4	68.21%	85.74%	76.81%	84.57%	58.13%	5.93%	21.13%	8.47%
5	70.20%	81.93%	88.96%	87.89%	33.93%	16.20%	6.00%	6.53%
6	54.96%	71.19%	82.78%	76.20%	68.27%	36.33%	2.80%	17.87%
7	75.94%	74.63%	81.40%	80.75%	14.53%	17.20%	5.07%	3.93%
8	25.17%	93.43%	90.72%	91.97%	96.67%	0.87%	5.53%	3.47%
9	59.65%	78.29%	80.90%	74.05%	66.53%	17.13%	12.47%	34.60%
Average (mean±std)	60.79% ±14.98%	83.06% ±8.12%	83.38% ±6.93%	84.70% ±7.24%	51.33% ±24.97%	12.37% ±10.82%	10.89% ±10.33%	9.25% ±9.99%

Table 6.2: Efficiency analysis — Average number of interactions required to reach the reward threshold as a percentage of total interactions (PI) and the number of adolescents achieving identified reward thresholds.

Translation Function	**Reward Threshold**			
	25%	**50%**	**70%**	**80%**
Linear	64.55% (10)	71.91% (9)	80.28% (5)	None
Quadratic	15.65% (10)	19.91% (10)	26.71% (10)	41.37% (8)
Proportional Quadratic	38.75% (10)	45.79% (10)	52.51% (10)	63.44% (9)
Exponential	16.96% (10)	21.22% (10)	25.97% (10)	34.90% (10)

identified reward thresholds and the number of subjects reaching the target threshold. The linear action space was unable to reach the 80% reward threshold and also took more PI to cross lower reward thresholds. The quadratic and proportional-quadratic functions also performed better compared to the linear action space.

The structure of the reward graphs (Figure 6.4) for the exponential and quadratic action spaces gave evidence of steady learning and better convergence for all subjects. The reward graphs of adolescent 3 and 5 clearly indicated unsteady learning for the linear action space where sudden large reward fluctuations are observed during training. The linear action space only converged to sub-optimal reward levels in some subjects (Adolescent 1, 6, 9), while it failed to achieve convergence in others (Adolescent 0, 8) within the target number of training interactions. However, it is expected that the linear action space-based RL algorithm might converge if the number of training interactions are increased. Meanwhile, all the proposed non-linear action spaces were able to converge within the target training interactions. The subjects who achieved convergence under the linear action space only converged to a sub-optimal level which indicates that increasing training interactions might not be beneficial for these subjects.

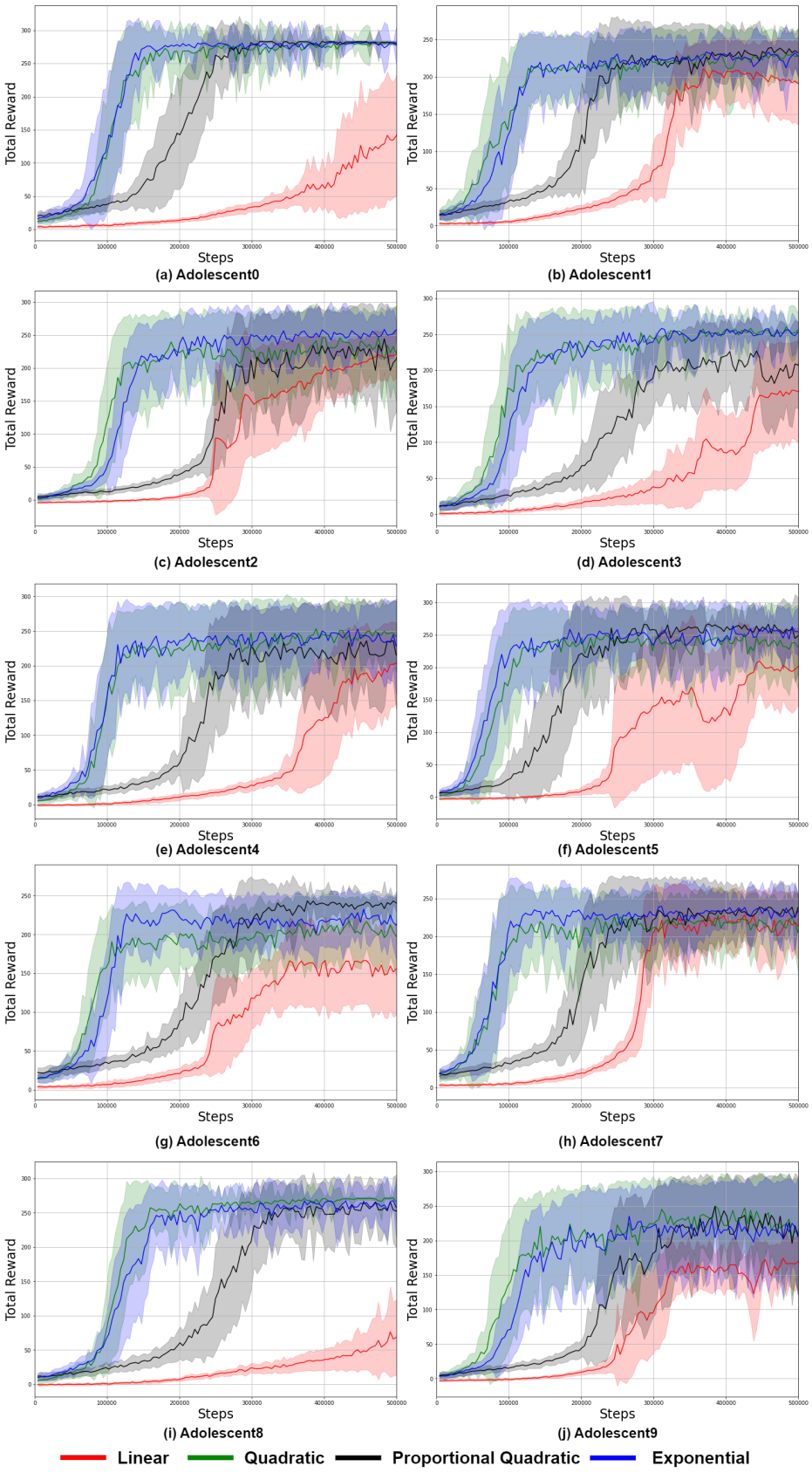


Figure 6.4: Comparison of candidate action spaces during training. The mean and standard deviation of the total testing reward (3 random seeds x 20 testing scenarios) achieved for each candidate action space is presented against 500,000 learning interactions.

6.4 Discussion

The application of RL algorithms to problems with large continuous action spaces is challenging and currently being tackled through the design of efficient exploration algorithms [Lazaric et al., 2007] and irrelevant/redundant action elimination [Zahavy et al., 2018]. The common practice in continuous control RL tasks present in OpenAI Gym [Brockman et al., 2016], DeepMind Control Suite [Tassa et al., 2018], and MuJoCo physics environments [Todorov et al., 2012] is to use a linear action space. Inspired by clinical treatment methods for T1D, non-linear continuous action space representations were introduced to tackle the challenge of the large, continuous, and non-uniform insulin action space. The results demonstrated superior performance of the non-linear action spaces compared to the standard linear one with enhanced learning efficiency, smoother convergence and improved performance of glucose control. The exponential action space performed the best out of the proposed non-linear action spaces. Hence, the exponential action space was used in the RL algorithms designed in Chapters 4 and 5. Based on the acquired knowledge through the literature survey presented in Chapter 3, this is the first attempt to explore action space representations to tackle complexities in large continuous action spaces associated with glucose regulation in T1D.

The learning efficiency achieved in the proposed approach indicates efficient exploration by the RL algorithm. This also reduces the computational time requirements for training, which is very valuable in the design and development phase of RL algorithms for glucose control, as often multiple iterations of designs are explored and tested. Increasing the complexity in the RL algorithmic architecture or the simulation protocol is expected to result in increased compute times until convergence is achieved. Hence, faster learning can become not only desirable but also vital for the experimental design of future RL algorithms for glucose regulation.

The successful real-world application of RL-based APS would require online continual learning to adapt the control strategy based on biological variability (e.g., ageing, hormonal disturbances) of the user. Hence, the steady learning observed in the exponential and quadratic action spaces is vital to ensure safety by avoiding sudden excessive changes. The proposed approach illustrated favourable characteristics in this regard, while further future research is required to evaluate the operation in an online continual learning setting. This approach can also be applied to other medical applications with similar action space properties. The application of propofol dosing in general anaesthesia is such an application where a non-uniform action distribution is observed for which RL approaches are currently being explored [Vajapey, 2019].

Chapter 7

Multi-input Artificial Pancreas Systems (MAPS)

In previous chapters, RL was explored to address technical and application-specific challenges related to glucose control. This chapter explores the potential of integrating additional physiological signals to address identified challenges and design MAPS. The chapter is divided into two sections. The first section presents a systematic literature review to identify additional physiological signals suitable for glucose control. Based on the review findings, the following section analyses invasive physiological signals in the context of exercise. Exercise is a challenging activity affecting glucose control due to the rapidly changing insulin needs. Relationships between invasive physiological signals and different types of exercise are explored to assist glucose control and design MAPS.

7.1 Background

The biological control mechanism within the body to regulate glucose is immensely complex. The β-cells which control the glucose level rely on multiple biochemical chain reactions [de la Vega et al., 2011]. In addition, multiple hormones closely participate in glucose regulation and control-related functions such as satiety, digestion, and growth [Teixeira and Malin, 2008]. The central nervous system is also involved in the process by modulating the responses of endocrinal and other tissues according to factors such as the circadian rhythm [Teixeira and Malin, 2008]. Hence, it is clear that the glucoregulatory system is complex and integrated with many processes. However, in current APS, only limited information is used, namely glucose measurements, administered insulin, and meals.

Integrating additional external inputs to APS can help overcome the above limitations by providing more information for better decision-making. In addition, the additional information could support the automatic identification of events such as meals, exercise, sleep, stress, and other biological variations which affect the glucose profile. This would be beneficial to reduce the risk of T1D complications. Detection of these activities would also help improve glucose control performance by minimising the limitations arising from glucose sensor/insulin action delays. Using additional

inputs would also help reduce the cognitive burden through less user interaction leading to a better quality of life.

The rapid development of wearable sensor technologies can be identified as a strong motivator towards **Multi-input Artificial Pancreas Systems (MAPS)** (Figure 7.1). In previous studies Kudva et al. [2014] analysed the clinical importance of additional physiological signals, while Cinar [2017] and Patek [2019] analysed the current limitations of APS to approach MAPS. However, evaluating the potential improvement and additional device burden arising through these systems is essential.

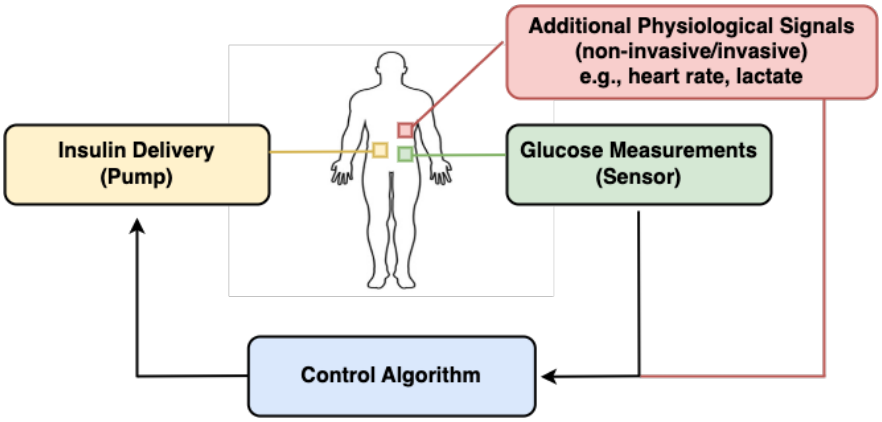


Figure 7.1: Multi-input artificial pancreas system.

7.2 Systematic Literature Review: Integrating Multiple Inputs Into APS

7.2.1 Introduction

MAPS seek to integrate different types of additional physiological inputs to improve the performance of glucose regulation. Previous studies have explored the clinical and technical importance of MAPS [Kudva et al., 2014; Cinar, 2017; Patek, 2019]. However, these studies present a general overview and do not explore the detailed technical aspects of MAPS development. Therefore, a systematic review was conducted to identify and understand the current state of MAPS by addressing the following research questions:

- What type of additional inputs have been utilised in designing APS, and why have they been integrated?
- What type of controller architectures and algorithms has been used in developing MAPS?
- What validation methodologies are carried out for evaluating the performance of the proposed systems?

Identifying the rationale and evaluating the importance of different types of additional inputs is a must to understand their potential value and effectiveness. The control architectures and methods used to combine the additional inputs must be explored to understand how they have been integrated into existing control algorithms. Developing control algorithms is challenging since it requires simulations or clinical trials for testing. Sound methodologies and simulations must be carried out due to the inherent risk in medical devices. Furthermore, the development of such systems needs to be appropriately validated before human testing. The development and validation methodologies are expected to be complex with the addition of various inputs compared to conventional APS. Hence, alongside the used control solutions, it is valuable to identify the different simulation models and experimental procedures carried out in previous research related to MAPS. This survey is motivated towards addressing these aspects and understanding the current state of MAPS.

7.2.2 Methodology

The literature survey was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework [PRISMA, 2020]. Scopus, IEEE Xplore (to capture engineering studies on APS development, including multiple-input scenarios), and PubMed (to capture APS clinical studies conducted corresponding to the identified engineering approaches) databases were searched for the period between January 1, 2005, and February 10, 2020. The survey focused on analysing and summarising the different input sources, in addition to glucose measurements, integrated into MAPS; control algorithms; architectures; and the validation methodologies used. The clinical transition of the identified studies was also considered to obtain a complete picture of the current state of progress of MAPS developments. The study selection process was carried out in 4 steps (Figure 7.2).

1. **Identification phase**: Different strategies were explored to develop a suitable search query with the help of a research librarian (Appendix B.1). Upon conclusion, a broad search query was developed to identify all papers related to control of APS. The search query was not restricted further to ensure that all studies related to MAPS were captured. For the Scopus database, the search was restricted to articles and conference proceedings related to the subject areas of Engineering, Computer Science, Mathematics, Decision Sciences, and Multidisciplinary. The subject area restriction was not possible in PubMed; thus, the query was adjusted to exclude review articles and only include research related to humans. No additional filtering was carried out in the IEEE Xplore database due to its specific focus on Computer Science & Electrical Engineering. The search queries are presented in Table 7.1. Additional records were identified through following references in selected studies.

2. **Screening phase**: The identified papers from the database search and other resources were first analysed to remove duplicates. Then, titles and abstracts of the remaining papers were screened to exclude papers on animal studies,

Table 7.1: Search queries.

Database	Search Strategy	Results
Scopus	(((close* AND loop) AND (diabet* OR t1d)) OR artificial W/2 pancreas) AND (control*); (Filter: Subject Area & Article Type)	668
PubMed	(((close* AND loop) AND (diabet* OR t1d)) OR artificial W/2 pancreas) AND (control*) NOT (review*); (Filter: Humans)	393
IEEEXplore	(((close* AND loop) AND (diabet* OR t1d)) OR "artificial pancreas") AND control*	327

glucose sensor development and errors analysis, insulin pumps, other aspects of APS (e.g. safety, user experience, psychosocial aspects), physiological modelling, studies related to the chemical, biological, and medical aspects of APS design, studies focussing on developing sub modules for APS (e.g. glucose level estimation, meal detection), studies without additional input signals and other irrelevant studies.

3. **Eligibility phase**: The remaining full papers were analysed and included in the study if the following selection criteria were met: (1) a control algorithm or architecture is present, (2) additional external inputs are used for the controller design (i.e., in addition to CGM measurements, and the additional inputs are not control inputs, such as the coinfusion of glucagon), and (3) a validation is conducted *in-silico* or *in-vivo* (in humans). These criteria were formulated to encompass the 3 main verticals of the survey to understand and summarise the use of additional wearables in APS design, APS development technologies, and validation methodologies.

4. **Inclusion phase**: Finally, the selected studies (N=17) were categorised into two groups. The main studies which introduce different unique MAPS (11/17, 65%) and their associated clinical studies (6/17, 35%). It is important to note that some of the main studies also included clinical trial results (3/11, 27%). This separation was required to avoid the duplication of similar APS and ensure the overall analysis of the identified technical criteria of the unique MAPS studies. The main studies were thoroughly analysed based on the additional inputs used, APS controller design, and validation methodologies which is presented in detail in the next section. The clinical studies were analysed in order to discuss the feasibility of MAPS.

Quality Assessment: A quality assessment of the selected 17 studies was carried out using the CASP (Critical Appraisal Skills Programme) tool [CASP, 2020] (Appendix B.2). The main issues highlighted by the assessment were the difficulty in ascertaining the risk of bias in data collection/simulation and the failure to provide

sufficient information regarding the ethics approvals.

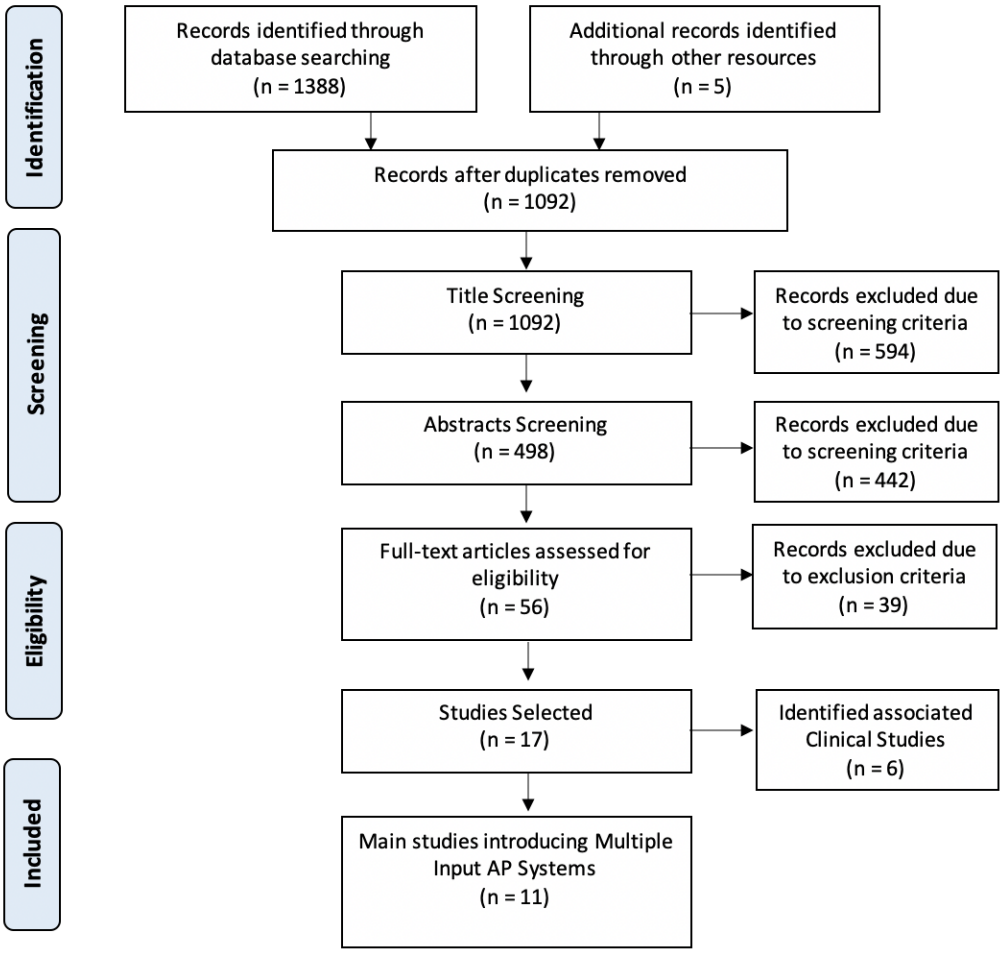


Figure 7.2: Study selection & identification flow chart.

7.2.3 Results

Non-invasive & Invasive Inputs. The types of additional inputs integrated or proposed to be integrated into APS that were identified can be categorised as (1) non-invasive inputs captured through wearable devices and (2) invasive inputs of substances measured in body fluids. Most (9/11, 82%) of the selected studies focused on using non-invasive wearable input for MAPS development. Electrocardiogram (ECG), HR, accelerometers, skin resistance, energy expenditure (EE), and galvanic skin response (GSR) were identified as the non-invasive sensory inputs, and lactate and adrenaline was identified as invasive inputs.

Table 7.2: Types of additional integrated inputs.

Type	Additional Input
Invasive	Lactate, Adrenaline
Non-invasive	Heart Rate, Accelerometer, ECG, Skin Resistance, Energy Expenditure, Galvanic Skin Response

The identified studies used the additional inputs mainly for exercise detection, hypoglycaemia prediction, and stress detection. A large portion of the studies focused on exercise detection. They mainly used HR, accelerometer, EE (also referred to as metabolic equivalent [MET]), and lactate for exercise detection. Turksoy et al. [2013a,d] and Hajizadeh et al. [2019] used the readily available EE data from wearables, whereas Jacobs et al. [2015] and Resalat et al. [2019b] used a regression model introduced by Zakeri et al. [2008] to convert HR and accelerometer data to calculate MET. Stenerson et al. [2014] used HR and accelerometer data, and DeBoer et al. [2017] used HR data to identify exercise through predefined threshold values. Quiroz and Femat [2010] and Quiroz et al. [2011] simulated the effects of lactate for exercise detection.

Hypoglycaemia prediction, using additional physiological signals, was the next popular approach to MAPS design. Predicting hypoglycaemia in advance helps mitigate the glucose-sensing delays. Khan et al. [2013] and Qaisar et al. [2012] used HR, ECG (QT interval), and skin resistance for hypoglycaemia detection. They identified the QT interval as the most prominent input, and skin resistance as the least important input in hypoglycaemia prediction. Turksoy et al. [2013c] used EE and GSR to develop a module for hypoglycaemia prediction. Quiroz and Femat [2010] and Quiroz et al. [2011] used adrenaline for hypoglycaemia detection. Stress detection was identified as another important aspect for MAPS design, where Turksoy et al. [2013a,d] focused on using GSR signals. A summary of the identified inputs and their target limitations in existing systems is presented in Figure 7.3

Control Architectures. The additional inputs identified in the previous section have been integrated with the control algorithm based on different architectures: (1) switching the controller mode through activity detection, (2) directly incorporating in the controller for decision-making, and (3) developing intermediate modules for the controller (e.g., hypoglycaemia and meal detection) (Table 7.4). These architectures are based on a variety of control algorithms (Figure 7.4), where adaptive model-based methods have been used the most in 37% of the studies.

Stenerson et al. [2014], DeBoer et al. [2017], and Jacobs et al. [2015] focused on switching the mode of the controller based on detected activity. HR and accelerometer input were used in this approach, where the controller mode was changed through adjusting parameters and thresholds within the controller. Stenerson et al. [2014] paused their predictive low-glucose suspend algorithm, and DeBoer et al.

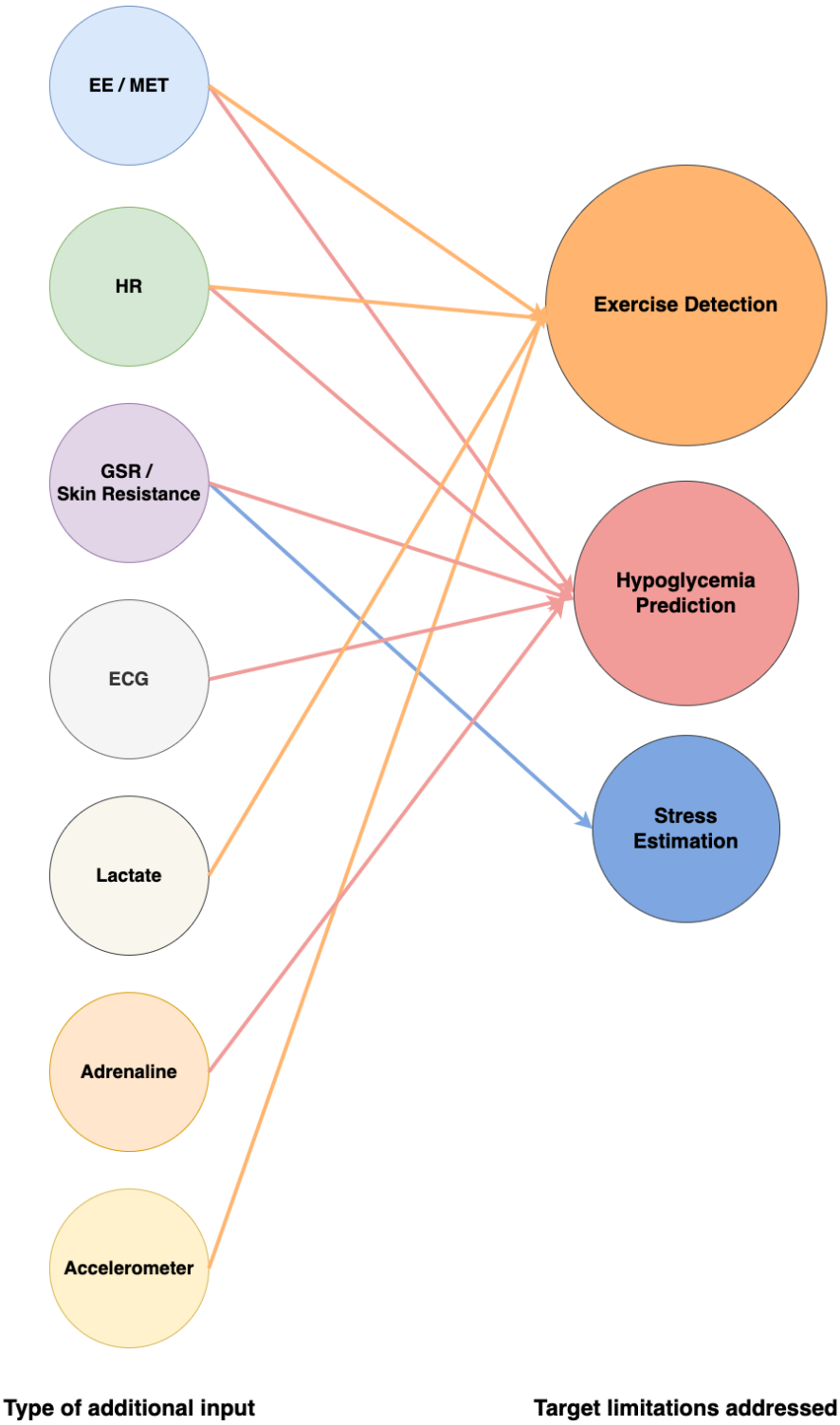


Figure 7.3: Additional inputs used in MAPS design and their main focus aspects.

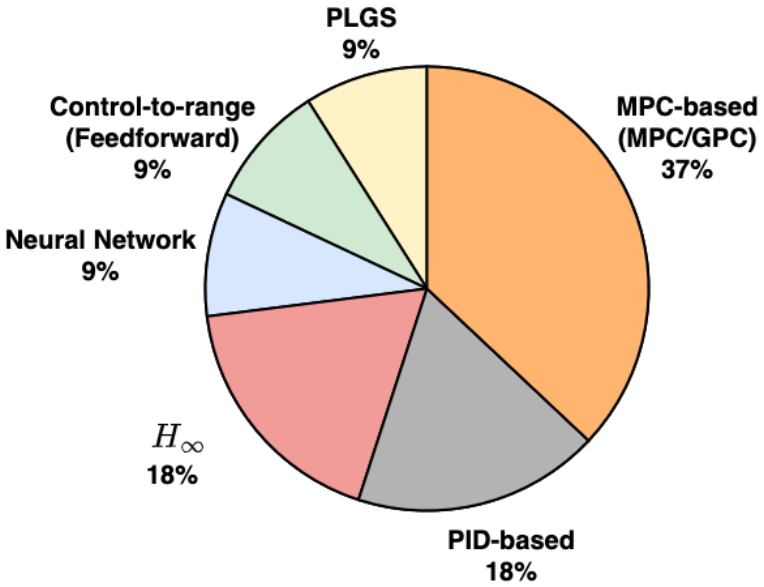


Figure 7.4: Types of controllers used by previous studies. Acronyms: GPC - Generalised Predictive Control, PLGS - Predictive Low Glucose Suspend, PID - Proportional Integral Derivative, MPC - Model Predictive Control.

[2017] adjusted the hypoglycaemia risk threshold in their control-to-range controller when exercise was detected. Jacobs et al. [2015] used a fading memory proportional-derivative dual-hormone controller that, upon the detection of exercise, carried out dosing of insulin and glucagon based on a set of static rules.

In addition to activity detection, additional inputs may contain valuable information related to glucose regulation. Hence, studies have focused on directly integrating additional inputs for decision-making. Quiroz and Femat [2010], Quiroz et al. [2011], Resalat et al. [2019b], Turksoy et al. [2013a,d], and Hajizadeh et al. [2019] focused on direct integration of additional inputs in the controller design. Quiroz and Femat [2010] and Quiroz et al. [2011] directly integrated lactate and adrenaline input into their H_∞ control algorithm. Resalat et al. [2019b] developed a run-to-run MPC that used continuous MET data for exercise detection. Turksoy et al. [2013a] integrated EE and GSR into a generalised predictive controller by developing time series models using autoregressive moving average with external input, recursive least squares, and constrained optimisation techniques. They improved on their work by introducing a time-varying forgetting factor for the weighted recursive least squares algorithm and focusing on trajectory tracking [Turksoy et al., 2013d]. Hajizadeh et al. [2019] used recursive subspace identification techniques to develop an adaptive MPC controller incorporating MET input.

Designing submodules for APS has also been widely explored, where the focus has been on using the input to enhance insulin and glucagon infusion and to design safety mechanisms for APS. These submodules were mainly linked to identified limitations such as meal detection, activity detection, and hypoglycaemia detection.

Khan et al. [2013] and Qaisar et al. [2012] developed a hypoglycaemia-detection module using HR, ECG (QT Interval), and skin resistance. In addition to the main controller focusing on insulin infusion, a fuzzy logic fusion controller was introduced to infuse glucagon based on the identified signals during a hypoglycaemia event. Turksoy et al. [2018] performed a clinical trial where hypoglycaemia early alarm [Turksoy et al., 2013c], meal detection [Turksoy et al., 2015; Ramkissoon et al., 2018], hypoglycaemia prediction, and carbohydrate recommendation [Turksoy et al., 2016] modules were integrated into the final APS design. Hajizadeh et al. [2019] focused on plasma insulin concentration estimation and meal effect estimation modules in their research. Resalat et al. [2019b] proposed and evaluated an insulin sensitivity adaptation algorithm and an adaptive-learning postprandial hypoglycaemia prevention algorithm. Different safety modules have also been introduced, where Turksoy et al. [2013a] and DeBoer et al. [2017] focused on hypoglycaemia and hyperglycaemia safety, respectively, through insulin-on-board estimates. The development of submodules enhances the interpretability of the APS operation, which is essential in safety-critical applications. Most of the studies have used submodules in their controllers, both with switching the controller mode through activity detection and when additional inputs are directly integrated. Hence, designing submodules using additional inputs that target identified limitations in APS is likely to be beneficial to their further development.

Validation Methodologies & Study Outcomes. The designed APS have been validated through simulations and clinical studies (Table 7.3, 7.4, and 7.5). A variety of physiological models and tools have been used for simulations and different protocols used for clinical trials. An FDA-approved simulator is currently unavailable for testing MAPS. Two research groups have focused on developing their own multiple-input simulators [Resalat et al., 2019a; Rashid et al., 2019], which would be beneficial for the progress of MAPS development. MATLAB was used in most of the studies to conduct simulations. Quiroz and Femat [2010] and Quiroz et al. [2011] simulated the use of invasive inputs based on the Sorenson model [Sorensen, 1985], the Bergman minimal model [Bergman et al., 1979], the glucose–adrenaline relationship discussed in the study by Schultes et al. [2007], and the glucose–lactate relationship discussed in the study by Kreisman et al. [2001]. Khan et al. [2013] and Qaisar et al. [2012] also used the Bergman minimal model, as well as simulated meals, ECG, and subcutaneous delays. Jacobs et al. [2015] used the Hovorka insulin pharmacodynamics model [Hovorka et al., 2004], the insulin pharmacokinetics model by Wilinska et al. [2005], the glucagon pharmacokinetics model by Lv et al. [2013], the glucagon pharmacodynamics model by Bakhtiani et al. [2015], and the exercise model by Hernández-Ordoñez and Campos-Delgado [2008] for their simulation. Resalat et al. [2019b] and Hajizadeh et al. [2019] conducted their simulations based on simulators developed by their own research groups [Resalat et al., 2019a; Rashid et al., 2019].

Breton et al. [2014], Jacobs et al. [2015, 2016], and DeBoer et al. [2017] carried out a clinical trial to evaluate their switching mode controller after obtaining FDA

and institutional review board approvals. Breton et al. [2014] and DeBoer et al. [2017] reported a reduction in hypoglycaemic events in adolescents and adults, respectively, using HR in an activity-augmented control architecture. Jacobs et al. [2016] also achieved a reduction in time spent in hypoglycaemia, but there was an increase in the time spent in hyperglycaemia when the exercise-augmented control structure was used. Similar results were observed in the subsequent trial by Castle et al. [2018]. Overall, these randomised crossover trials were able to identify a reduction in hypoglycaemia when the activity-augmented control structure was used. It is important to note that activity-augmented APS design might be compromised during different types of exercises (high-intensity training and resistance exercise), which has not been explored.

Turksoy et al. [2013b, 2014, 2018] focused on having a medical expert to review each insulin dose before the application and obtained institutional review board approval. They focused on integrating continuous inputs (EE and GSR) into the controller and developing submodules and conducted clinical trials for evaluation. They succeeded in improving the time in target range (70-180 mg/dL) to 76.75% with the integration of different submodules into the APS. The identified clinical trials in this survey (Table 7.5) focused on either adolescents, young adults, or adults. The trials comprised both normal closed-loop trials and randomised crossover trials, which evaluated different treatment types and typically ranged in duration from 1 to 4 days. Further longitudinal studies will be beneficial to ascertain the effects of sensor noise and unanticipated dropouts that might arise from the additionally introduced sensors. It is important to conduct trials encompassing all age groups (children, adolescents, and adults) to evaluate the robustness of the controllers because different age groups have different insulin sensitivities, which affects the controller's accuracy.

Table 7.3: Summary of simulation models used.

Study	**Simulation Model**
Quiroz and Femat [2010]	Sorenson's Model [Sorensen, 1985], Bergman's Minimal Model [Bergman et al., 1979], Glucose – adrenaline relation - Schultes et al. [2007], and Glucose – lactate relationship - Kreisman et al. [2001].
Quiroz et al. [2011]	Sorenson's Model [Sorensen, 1985], Bergman's Minimal Model [Bergman et al., 1979], Glucose – adrenaline relation - Schultes et al. [2007], and Glucose – lactate relationship - Kreisman et al. [2001].
Khan et al. [2013]	Bergman's Minimal Model [Bergman et al., 1979].
Qaisar et al. [2012]	Bergman's Minimal Model [Bergman et al., 1979].
Stenerson et al. [2014]	Not Specified.
Jacobs et al. [2015]	Hovorka's insulin pharmacodynamics model [Hovorka et al., 2004], insulin pharmacokinetics model by Wilinska et al. [2005], glucagon pharmacokinetics model by Lv et al. [2013], glucagon pharmacodynamics model by Bakhtiani et al. [2015], and exercise model by Hernández-Ordoñez and Campos-Delgado [2008].
Resalat et al. [2019b]	Virtual patient population by Resalat et al. [2019a].
Hajizadeh et al. [2019]	mGIPsim simulator (2019) [Rashid et al., 2019].

Table 7.4: Summary of selected studies.

Study	Additional Inputs	Controller	Architecture	Validation
Quiroz and Femat [2010]	Lactate, Adrenaline	H_∞	Additional Inputs directly integrated.	Matlab Simulation
Quiroz et al. [2011]	Lactate, Adrenaline	H_∞	Additional Inputs directly integrated.	Matlab Simulation
Khan et al. [2013]	ECG, HR, Skin Resistance	PID	Fuzzy fusion controller to fuse the additional input to prompt glucagon infusion (Dual Hormone).	Matlab Simulation
Qaisar et al. [2012]	ECG, HR, Skin Resistance	Neural Network Predictive	Fuzzy fusion controller to fuse the additional input to prompt glucagon infusion (Dual Hormone).	Matlab Simulation
Stenerson et al. [2014]	HR, Accelerometer	PLGS	Additional Inputs used to switch between modes.	Simulator (unspecified)
DeBoer et al. [2017]	HR	Control to Range	Additional Inputs used to switch between modes (only basal rate is controlled).	Clinical Study
Jacobs et al. [2015]	EE (HR, Accelerometer used to calculate)	FMPD	Additional Inputs used to switch the controller to a different mode (Dual Hormone).	Simulation, Clinical Study
Resalat et al. [2019b]	MET (HR, Accelerometer)	Adaptive Run to Run MPC	Inputs used to calculate MET which is directly used by the Controller for decision making. Meal Data also provided to the controller.	Simulation
Turksoy et al. [2013a]	EE, GSR	GPC	Inputs integrated directly. ARMAX, Recursive Lease Square & Constrained Optimisation used.	Clinical Study
Turksoy et al. [2013d]	EE, GSR	GPC	Inputs integrated directly, Time varying forgetting factor for WRLS Algorithm & trajectory tracking.	Clinical Study
Hajizadeh et al. [2019]	EE(MET)	Adaptive MPC	Additional inputs integrated directly to the controller. Recursive subspace identification techniques, PIC, meal estimates also used as inputs to the controller.	Simulation

Table 7.5: Comparison of clinical trial results.

Study	Trial/Controller Setting	Results
Breton et al. [2014]	12 adults, randomized cross over trial, 24hour closed loop experiments each with exercise. Exercise detection using HR. (Meal bolus manually calculated)	TIR for APS with HR and without HR, Overall 81% Vs 75%), Exercise 91% vs 85%, Overnight 89% vs 84%. Using HR resulted in fewer hypoglycaemic events during exercise (none vs 2).
DeBoer et al. [2017]	18 adolescents, randomized cross over trial, 24hour closed loop experiments each with exercise. Exercise detection using HR. (Meal bolus manually calculated)	TIR for APS with HR and without HR. Overall 77% vs 74%, Exercise 96% vs 87%, Overnight 92% vs 84%. Small reduction in hypoglycaemic events (0.39 HR informed AP vs 0.50 without HR).
Jacobs et al. [2016]	21 adults, randomized crossover trial with, 22-hour experiments each with exercise. Exercise detection Algorithm triggered manually.	TIR with exercise detection 67%, without exercise detection 72%, SAP 68%. Time spend in hypoglycaemia (<3.9mmol/L) 0.3%, 3.1%, 0.8% respectively. Time spend in hyperglycaemia (<10mmol/L) 32%, 25%, 31% respectively.
Castle et al. [2018]	20 adults, randomized crossover trial, 4-day experiments each with exercise. Exercise Detection Algorithm triggered using wearable sensor in SH, DH controllers.	TIR overall single hormone 74.3%, dual hormone 72%, PLGS 65.2%, current care 63.1%. Time in Hypoglycaemia 2.8%, 1.3%, 2%, 3.1% respectively.
Turksoy et al. [2013a,b]	3 young adults, seven 32 or 60 hour closed loop experiments with exercise. Additional signals integrated continuously.	TIR 62%. (Overnight: 75.3%, Exercise: 55%, Glycemic closed loop: 56.1%)
Turksoy et al. [2013d]	3 young adults, 70-hour closed loop experiments with exercise. Additional signals integrated continuously.	TIR 46.5%.
Turksoy et al. [2014]	9 young adults, 2-day closed loop experiments with exercise. Additional signals integrated continuously.	TIR 58%.
Turksoy et al. [2018]	10 young adults, 18, 60-hour closed loop experiments with exercise. Additional signals integrated continuously, with sub modules.	TIR 69.9% for exercise and recovery periods, 76.75% overall performance.

7.2.4 Discussion

Principal Findings. This survey focussed on three main verticals, (1) identifying the types of additional input signals used, (2) analysing control architectures, and (3) exploring validation methodologies for MAPS. The majority of the identified inputs were non-invasive and captured through wearable devices. The effectiveness of invasive inputs was analysed through simulations, where lactate and adrenaline were explored for exercise and hypoglycaemia detection. EE (or MET) can be identified as the most frequent non-invasive additional input used in MAPS. Past studies have mainly focused on using non-invasive additional inputs for detecting exercise (HR, accelerometer, and EE), hypoglycaemia (ECG, HR, EE, and GSR), and stress (GSR). The technological advancements in wearable devices would benefit the development of MAPS. In future, additional sensors might also be valuable in capturing other physiological changes, such as illnesses, alcohol consumption, and seasonal variations.

Most studies (73%, 8/11) focused on augmenting single hormone APS compared to dual hormone systems. Identifying additional inputs which can be used to address current limitations and directly integrating those inputs with the controller has shown promise. Developing submodules based on these limitations and switching the controller's mode via activity detection can also be identified as effective approaches towards MAPS design. Validations were carried out in both *in-silico* and *in-vivo*. However, comparing the results is subjective due to the different physiological models used in the simulators and different protocols used in the clinical studies. Both quantitative and qualitative metrics were used to evaluate the effectiveness of the proposed systems. Some of these measures were the time in hypoglycaemia, normoglycaemia, hyperglycaemia ranges, and the number of hypoglycaemic events. HR, EE, GSR, and accelerometer-based studies have been evaluated through clinical trials mainly due to the easy access from wearable devices. The lack of an FDA-approved simulator for testing the identified additional input can be identified as a significant constraint regarding the development of MAPS.

In addition to the systematic literature review, a patent search was conducted on Google Patents from January 2005 - May 2021 to identify possible technological advancements, where two patents associated with MAPS were identified (Appendix B.3). Patent ID US8690820B2 [Cinar and Oruklu, 2014] presented a device where a glucose sensor and physiological status–monitoring system communicates with an automatic controller for glucose control. The controller also included a module to predict future glucose levels. Patent ID US10646650B2 [Cinar et al., 2020] introduced additional modules for recursive model identification of hypoglycaemia and hyperglycaemia early alert and alarm, plasma insulin concentration estimation, physical activity assessment, stress detection and assessment, sleep detection, and sensor and pump fault detection and diagnosis.

Feasibility of MAPS. Different additional inputs have been identified and used to address limitations identified in current generation APS. However, more signals and relationships need to be explored to address limitations such as meal and illness estimation. It is essential to quantify the improvement of the APS through the integration of additional input signals. The results of the proposed approaches can be analysed based on their clinical trials, which provides a fairer interpretation than the simulations. However, comparing trials is not straightforward due to the different protocols (meals and exercise) and the number of subjects present. The identified clinical trials improved the time in the normoglycaemia range and reduced hypoglycaemic events when additional inputs were utilised. The noise and instability associated with wearable sensors also need to be evaluated, as they could have a detrimental effect on the controllers. Precautionary measures should be set in place to ensure patient safety during such circumstances. It is also important to note that the real-world application of MAPS would be very complex. For example, a person with T1D might not wear additional wearables during sleep, requiring the controller to work in highly dynamic environments. Hence, evaluating such scenarios through simulations and conducting clinical studies for larger cohorts and trial durations is a must to ensure the systems' effectiveness.

The benefits of MAPS should outweigh the burden of external sensors. Hence, the device burden of MAPS needs to be explored with the consultation of people with T1D and health experience experts. People with T1D who choose not to conduct multiple daily blood glucose level tests and use multiple daily injections are currently compelled to use minimally invasive CSGM and CSII devices. This requires users to take the necessary steps to regularly change the sensors [Brew-Sam et al., 2021a; Iturralde et al., 2017; Lawton et al., 2018; Barnard et al., 2017; Rankin et al., 2018]. Hence, an additional invasive sensor might be considered burdensome. However, there exists the possibility of integrating additional sensors in currently used devices such as CSGM and CSII to avoid additional user burden. Previous studies have identified relationships between invasive inputs and T1D (e.g., ketone sensors to identify diabetic ketoacidosis [Lee et al., 2019]). Although rich relationships exist, progress is stunted because of the lack of sensors for carrying out continuous measurements. At present, real-time interstitial insulin sensors and ketone sensors are being developed [Teymourian et al., 2019]. These practical complexities and psychosocial aspects must be evaluated to develop effective MAPS.

7.3 Analysing the Effect of Invasive Physiological Signals During Exercise

7.3.1 Introduction

The systematic literature review in Section 7.2 highlighted the benefits of using additional input signals and identified that non-invasive wearable sensor inputs (e.g., HR, accelerometer data) have been extensively explored compared to invasive inputs. Hence, this section explores the effects of invasive physiological signals and their relationships related to the glucoregulatory system. Meals and exercise are two critical daily events that affect glucose homeostasis. In previous chapters, the effects of meals were considered using T1D simulators and control algorithms designed to mitigate the impact of meals on glucose control. However, due to the limitations of existing simulators, the effects of exercise cannot be simulated. To overcome this limitation, studies identified in the systematic literature review conducted clinical trials and designed custom simulation environments to evaluate the impact of exercise. In contrast, in this thesis, the relationships between invasive physiological signals and exercise are studied, aiming to develop an effective glucose control strategy.

Exercise is challenging to existing treatment due to the rapidly changing insulin needs related to different types of exercise, and often forward-planning is required by people with T1D [Paldus et al., 2022a]. The large fluctuations in glucose during exercise also exacerbate the effect of sensor delays [Zaharieva et al., 2019]. The effect of exercise varies based on the type of exercise, and its duration can exceed the period of exercise [Paldus et al., 2022a]. Exercise detection is a straightforward problem that can use wearable device data such as HR, accelerometer, and invasive biochemicals. Previous studies on wearable devices have extensively explored exercise detection [Stenerson et al., 2014; Iqbal et al., 2016; DeBoer et al., 2017]. However, practical limitations may affect wearable devices (e.g., the device not being properly fitted and errors in HR measurements due to darker skin tones). Integrating additional invasive sensors into existing CSII, CSGM devices may provide better measurements for glucose control without additional device burden [Lee et al., 2019]. Recent studies on T1D have explored relationships between exercise and blood hormonal and metabolite biomarkers such as catecholamines, lactate and ketones [Lee et al., 2019; Paldus et al., 2022b]. The benefits of blood biomarkers are not restricted to exercise-related glucose control. They can also be used for vital tasks such as identifying diabetic ketoacidosis to ensure patient safety [Lee et al., 2019]. Hence, in this section, biochemical data is analysed. Specifically, the focus is on identifying relationships between biochemicals and different exercise types (resistance exercise (RE), high-intensity exercise (HIE), and moderate-intensity exercise (MIE)). Identifying potential leading indicators of exercise is valuable to mitigate the effects of sensor delays. Furthermore, the hormonal and metabolite biomarkers additional data may provide further opportunities to improve existing APS.

7.3.2 Methodology

The analysis was conducted based on data collected by St. Vincent's Hospital, Melbourne and ethics approvals were obtained for the use of the data for this analysis (St. Vincent's Hospital, Melbourne, Human Research Ethics Committee Reference Number HREC 021/18) [Paldus et al., 2022b]. The dataset consisted of biochemical information from 30 participants based on a randomised crossover trial, where each participant underwent three different types of exercise: RE, HIE, and MIE. The biochemical data were collected every 20 minutes through blood measurements during the exercise trials, which included 1-hour pre-exercise, 40-minute exercise, and 5-hours post-exercise periods. The blood hormonal and metabolite biomarkers measures included glucose, insulin, growth hormone, cortisol, glucagon, noradrenaline, adrenaline, dopamine, ketones, and lactate.

Multivariate Transfer Entropy (MTE) [Lizier and Rubinov, 2012] method was used to identify causal relationships among the blood biomarkers. MTE is an information-theoretic approach based on the transfer entropy [Schreiber, 2000], which can be used to identify the existence of relationships, direction, and time lag of information flow between processes using a model-free analysis. Transfer entropy methods are applied widely in neuroscience to understand the information flow between neural circuits through multichannel brain recordings [Wibral et al., 2014]. Transfer entropy [Schreiber, 2000] and Granger causality [Granger, 1969] are popular approaches to capture relationships between processes. However, these approaches have limitations when analysing multiple processes as the generally used bivariate forms may infer spurious or redundant relationships (e.g., common cause and pathway effects) [Lizier and Rubinov, 2012]. Furthermore, even though multivariate forms of these approaches exist, they are computationally expensive due to their brute-force nature and suffer combinatorial explosions [Wollstadt et al., 2018]. The MTE approach aims to overcome these challenges by following an iterative greedy approach to inferring relationships using incremental conditional transfer entropy measurements.

The IDTxl Toolkit [Wollstadt et al., 2018] was used to conduct experiments, where the analysis first combined all exercise types to identify relationships between blood biomarkers during exercise. Next, individual exercise types (RE, HIE, MIE) were analysed to identify exercise-specific relationships. The relationships identified in the experiments were based on a statistical significance level of $p = 0.05$ with adjustments for multiple comparisons.

7.3.3 Results

The experiment conducted by combining all types of exercise identified blood lactate as a leading indicator that can be used to estimate other related biochemicals in advance during exercise. The combined exercise analysis and the individual RE and HIE exercise types observed a relationship between lactate to cortisol. The network graphs presented in Figure 7.5 summarises the identified relationships, where the circles represent the blood biomarkers, while the arrows represent existence, direction, and time lag of the information flow of the relationships. The time lag is presented

based on the sampling rate of the data (1 = 20 minutes). The relationship between lactate and cortisol has also been identified in previous work based on other analysis techniques [Paldus et al., 2022b].

A relationship from noradrenaline to lactate was identified only in RE. This would be valuable to distinguish RE from MIE and HIE types for glucose control as different exercise types have varying effects on overall glucose. Relationships between lactate - growth hormone and glucose-ketones, were also observed in RE. Relationships between lactate, noradrenaline, and cortisol were not observed in MIE, while the only relationship identified in HIE was lactate to cortisol.

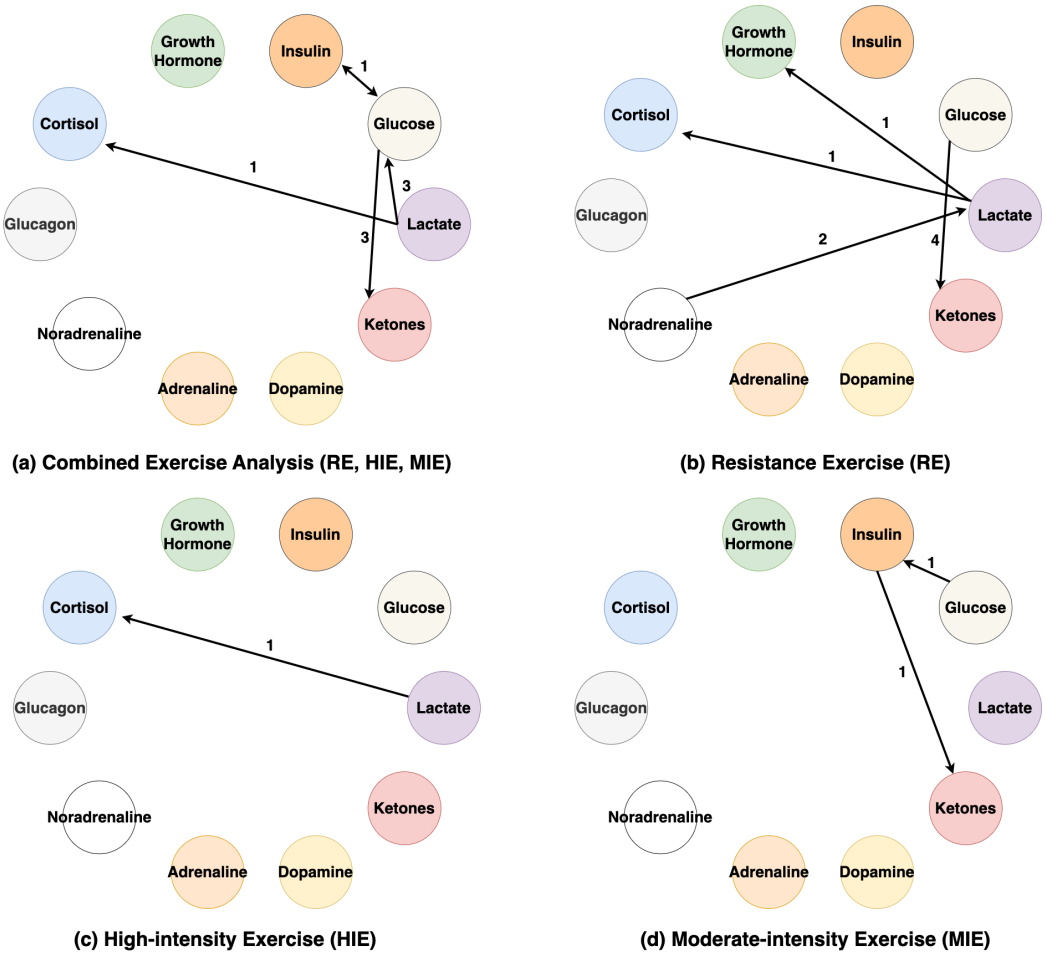


Figure 7.5: Information flow between blood hormonal and metabolite biomarkers. (a) Combined analysis for all exercise types, (b) resistance exercise, (c) high intensity exercise, and (d) moderate intensity exercise. The circles represent the blood biomarkers, while the arrows represent the existence, direction, and time lag of the information flow. The time lag is presented based on the sampling rate of the data (1 = 20 minutes).

7.3.4 Discussion

Knowledge of inter-relations between physiological signals can deepen the understanding of the glucoregulatory system and the design of personalised modelling and treatment solutions. The MTE analysis validated lactate as an important metabolite biomarker, which provides information regarding exercise. The systematic literature review also highlighted the importance of the use of lactate [Quiroz and Femat, 2010; Quiroz et al., 2011]. Differentiating different exercise types is important for selecting appropriate treatment as they have a varying effect on glucose [Paldus et al., 2022b]. Noradrenaline was observed to be valuable in differentiating RE from HIE and MIE types. Using lactate and noradrenaline as additional signals for APS can potentially improve glucose regulation during challenging exercise conditions. However, future research in many fields is required to use these blood biomarkers in a MAPS setting. Biochemical sensors capable of continuously measuring lactate and noradrenaline need to be developed. In contrast, this analysis collected data by conducting blood tests. These biochemical sensors should be integrated into existing CSGM devices, which measure glucose in subcutaneous interstitial fluid, to reduce the additional burden.

The analysis can be further improved by addressing several identified limitations. People with T1D who participated in the study were on a closed-loop treatment strategy. Hence, the treatment affects glucose and insulin behaviour, and true variation due to exercise cannot be identified. Furthermore, the sample size of the study was limited to 30 participants, while the biochemical data were sampled every 20 minutes, which affected the accuracy of the time lags calculated in the MTE method. However, conducting additional blood tests to collect data more frequently is practically infeasible. Hence, future analysis should progress with sensor development for near-continuous measurements. Nevertheless, this study highlights the potential of MAPS.

Chapter 8

Future of Artificial Pancreas Systems Development

Integrating AI into medicine has gained much attention due to the potential increase in the effectiveness and efficiency of medical practices. The screening of radiological images is a leading application that is currently being explored [Cabitza et al., 2021]. The ***human-AI collaboration*** plays a pivotal role in this aspect, where proper processes and protocols need to be explored [Cabitza et al., 2021]. For decades, the field of AI has focused more on competition between ***humans vs machines***. The former world chess champion Garry Kasparov has been at the forefront of this competition. He identifies the importance of collaboration and has presented an adage commonly referred to as the *Kasparov's Law*.

Kasparov's Law [Kasparov, 2017, p.236]:

*"**Weak human** + **machine** + **better process** was superior to a **strong computer** alone, and more remarkably, superior to a **strong human** + **machine** + **inferior process**."*

This highlights the importance of the less explored human-AI collaboration and the impact of having better processes and protocols for success. This adage was recently tested in a study by Cabitza et al. [2021], which involved radiologists and AI models to detect knee lesions through MRI (magnetic resonance imaging). The study verified the importance of better interaction protocols and the potential impact of human-AI collaboration.

This chapter discusses essential characteristics of APS development by first discussing the importance of collaboration between various stakeholders associated with the research. Next, two software tools named CAPSML and GluCoEnv are introduced to improve the communication barrier and research and development processes in RL-based APS. Finally, the future of APS are discussed by highlighting important characteristics and processes related to safety, privacy, ethical considerations, and clinical trials, which need to be further explored.

8.1 Background: Trends in APS Development

The concept of co-creation is a collaborative initiative between multiple stakeholders to create value through designing products and services [Makhni, 2017]. Co-creation practices are widely adopted in technology, education, retail, and financial services industries, while adoption in healthcare has been low [Makhni, 2017]. Traditionally, in medical practice, the patient plays a passive role, while the information flow and advice is from the healthcare professional who has the patient's best interest [Dinneen and McMorrow, 2022]. However, people with chronic conditions such as T1D and their carers become as knowledgeable and competent in the day-to-day self management [Dinneen and McMorrow, 2022]. This patient expertise fueled by the slow pace in the development of APS, have resulted in several open-source initiatives to design do-it-yourself APS [Kesavadev et al., 2020]. These open-source movements, such as #WeAreNotWaiting [Kesavadev et al., 2020] call attention to the timely need for APS solutions, highlights the value of co-creation, as well as the importance of proper regulatory frameworks.

> *"I do understand the regulatory concern, but you have to understand that, when done correctly, and I get that's a very big question mark, ... this technology is literally life changing.." - Parent of a person with T1D [Carolyn et al., 2020, p.1].*

The rapid advances of AI in health and medicine highlight the timely need for computer science researchers and engineers to be part of the co-creation process of APS development. The active involvement of all stakeholders in the AI implementation process would help them understand the benefits and limitations of AI [He et al., 2019]. Several studies conducted in T1D identify the importance of co-creation across disciplines, where people with T1D and their caregivers, clinicians, health experience experts, and researchers work actively towards designing and developing APS [Desborough et al., 2022; Brew-Sam et al., 2023]. The better understanding gained from a co-creation process would help explore solutions for challenges related to regulations, ethics, and accountability aspects of RL-based APS.

8.2 Software Tools for APS Development

This section presents the engineering contributions of this thesis by introducing an online demonstration tool named CAPSML and a GPU-based T1D simulator called GluCoEnv.

8.2.1 CAPSML: An Online Demonstration Tool

Technologies based on AI are increasingly being explored in clinical applications and have demonstrated synergistic effects when people and AI work together [He et al., 2019]. Hence, it is essential to highlight that excellence in designing AI for insulin delivery requires close collaboration of people and AI. However, the communication between AI and clinical experts across disciplines, professions, and areas

of expertise (including people with lived experience of T1D as research partners) is often lacking in existing research [Desborough et al., 2022]. This results in AI systems being designed which may not reflect the clinical objectives and expectations of people with T1D. The inherent complexity of AI algorithms further exacerbates the situation where the systems' understanding and explainability, operation, outcomes, data analysis mechanisms, and ML methods are low, leading to delays and failures in adoption. For example, the design of the reward function of an RL algorithm or the cost function of any ML algorithm requires expert domain knowledge (obtained, e.g., from clinicians and people with lived experience of T1D), which is vital. Hence, clinical and lived experience experts must be actively involved in the development process in a dynamic collaboration where they would enhance their knowledge and understanding of these novel technologies and provide their expertise and feedback to optimise the technologies further. AI researchers and systems developers acknowledging and appreciating this crucial domain expertise is an equally important priority. Furthermore, in Section 5.3, existing constraints in benchmarking and evaluating APS proposed in previous work were identified. The lack of proper benchmarks and open-source benchmarking platforms is a major limitation in current APS research, affecting proper communication and research translation.

Therefore, this thesis introduces an online tool to facilitate collaboration between clinical experts, lived experience experts and AI system developers toward developing APS. The tool named "CAPSML" offers the opportunity to interact with an AI system that simulates glucose regulation, to better understand its operations, improve the AI system, and propose improvements through clinical and lived experience feedback. In addition, the tool has the potential to provide free and accessible benchmarking, where researchers can submit algorithms securely and privately, which is absent in the diabetes technology community. "CAPSML" is publicly available online and can be accessed at https://capsml.com/ (Figure 8.1). The tool requires no specialised hardware and can be accessed through a desktop or mobile device. It allows users to experiment with AI algorithms designed for APS in T1D. Its demonstration video can be accessed at https://youtu.be/JO5MkPCuqCw.

The CAPSML tool has been designed based on the open-source Simglucose simulator [Xie, 2018] for conducting simulations. The tool contains implementations of the novel G2P2C algorithm, the state-of-the-art RL algorithms A2C, PPO, and SAC, and basal-bolus clinical treatment approaches BBI and BBHE and includes functionalities to define protocols, as well as to compare and analyse glucose regulation performance. First, the tool provides the capability for its user to define various experiment scenarios, namely, select *in-silico* T1D subjects from different age cohorts and customise meal protocols, that is, meals (breakfast, lunch, and dinner) with variable carbohydrate content and occurrence time (Figure 8.2). Second, the tool eases and enhances abilities to compare and analyse glucose regulation performance by examining the glucose/insulin trajectories and clinical metrics. The user has the ability to compare the glucose regulation performance of implemented AI systems against existing clinical treatment based on a basal-bolus strategy (Section 4.2.3).

The proposed CAPSML tool welcomes clinical and lived experience experts to

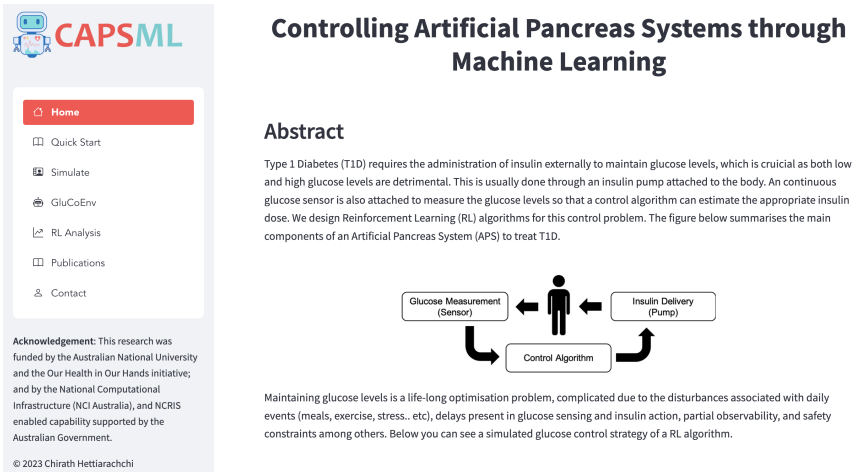


Figure 8.1: CAPSML - An online demonstration tool.

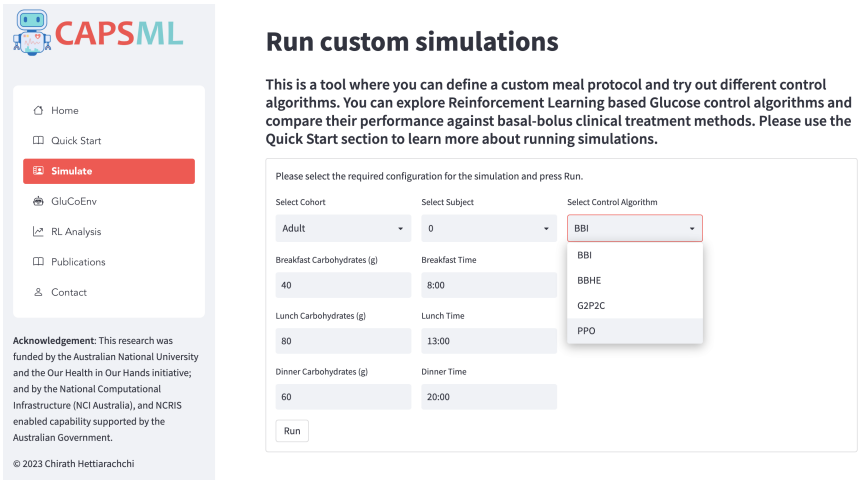


Figure 8.2: CAPSML - Running custom simulations.

experiment with an AI system and run custom glucose simulations. It also provides the opportunity to investigate and learn about the AI system, which is valuable for gaining insights into the treatment strategy learned by the intelligent algorithm. By conducting custom simulations, the strengths and weaknesses of the system can be identified. This is expected to be valuable for both the AI system and clinical experts as the AI system could be continuously improved and fine-tuned based on the knowledge bases of clinicians and people with lived experience of T1D; human-machine collaboration would be a first step toward improving the explainability of AI systems and accelerating their possible adoption in clinical practice. The proposed tool incentivises researchers to submit their AI algorithms for comparison and evaluation. If successfully used at scale, these systematic submissions could address the currently limited benchmarking of different AI algorithms and systems in T1D.

8.2.2 GluCoEnv: A GPU-based T1D Simulator

The development of RL-based APS requires significant effort to explore suitable problem formulations, development of RL algorithms, tuning of hyperparameters, and testing on complete cohorts provided in simulators. Hence, the process requires significant compute and time to run the experiments. This is a limitation in RL-based APS development.

Similar bottlenecks in the development pipeline exist in many other RL problems. Many recent research efforts have focussed on methods to mitigate them [Makoviychuk et al., 2021; Lan et al., 2021; Freeman et al., 2021; Weng et al., 2022]. A major bottleneck is the latency in communication added between the Central Processing Unit (CPU) and the Graphical Processing Unit (GPU) of the computer. This is because the simulation environment is often implemented on the CPU, while the RL algorithm based on neural networks is more efficient and suited for GPUs. During each environment step the communication between the simulator and the RL agent is essential, which is challenging when thousands or millions of training steps are required. Hence, previous work has focused on implementing simulators to run on GPUs [Makoviychuk et al., 2021; Lan et al., 2021; Freeman et al., 2021]. RL algorithms can also benefit from parallelism, where parallel agents are used during training to collect experiences. Hence, previous work has focused on designing vectorised parallel environments [Freeman et al., 2021; Weng et al., 2022].

In this thesis, following engineering best practices identified in previous work on RL focused simulators, a tool named GluCoEnv (Glucose Control Environment) for T1D simulations, is introduced. GluCoEnv is developed based on the Simglucose [Xie, 2018] simulator and aims to address the bottlenecks identified in the RL development pipeline. Therefore, GluCoEnv was designed to run entirely on a GPU and provide parallel environments to improve the throughput of experiments. GluCoEnv is open-source under the MIT license and available at https://github.com/chirathyh/GluCoEnv. The documentation of using GluCoEnv is available at https://glucoenv.readthedocs.io/en/latest/index.html. The tool is currently in active development. As part of the tool, benchmark scenarios are proposed, providing added flexibility to explore different problem formulations. The development of the UVA/PADOVA 2008 simulator [Kovatchev et al., 2009] has greatly impacted the development of control algorithms for APS, and the Simglucose 2018 [Xie, 2018] made it possible for the application of RL in this domain. It is expected that the engineering contributions of GluCoEnv would further strengthen the RL-APS research by enhancing the development pipeline. It also has the potential to attract more RL researchers to explore solutions for the glucose control problem and contribute towards developing real-world RL systems.

8.3 Future Directions and Recommendations for RL-based APS

8.3.1 Safety and Explainability

In this work a modular RL algorithm named G2P2C for glucose control was introduced which improved the performance and safety compared to the state-of-the-art PPO algorithm. The G2P2C algorithm was designed by considering application-specific constraints. The algorithm improved safety by introducing a planning mechanism. However, in future work the safety of G2P2C may be further improved by leveraging the learned glucose dynamics model to design additional safety modules. For example, the estimated future glucose may be used to forecast future hypoglycaemia and hyperglycaemia. In existing treatment methods such modules exist to inform the person with T1D so that the person can intervene and take necessary precautionary measures.

The glucose estimation module may be further improved to include the latest neural network architectures. This thesis used LSTM and dense networks in the glucose prediction module. Temporal and dynamic relationships between the glucose and insulin features used are essential for the glucose prediction task [Cui et al., 2021]. Latest deep neural network architectures, such as Transformers (2017), can learn such relationships through their self-attention mechanism [Vaswani et al., 2017]. During the thesis, a collaboration study explored the application of recurrent self-attention networks and transfer learning for glucose prediction and achieved the state-of-the-art performance on the standard OhiohT1DM benchmark [Marling and Bunescu, 2020; Cui et al., 2021]. Hence, G2P2C has the potential to improve by exploring the latest neural network architectures to augment the novel optimisation phases proposed in G2P2C.

Explainable AI (XAI) is an emerging field crucial for the practical deployment of AI models [Arrieta et al., 2020]. *Explainability* is an active characteristic of a model, where the model carries out an action or procedure with the intent of clarifying its function [Arrieta et al., 2020]. Explainability is very important in healthcare applications, particularly in RL-based APS, as the algorithm makes decisions frequently (e.g., every 5 minutes) compared to common uses of AI with one-time predictions [Gottesman et al., 2019]. As identified in Chapters 4 and 5, the RL-based APS decisions are more complex than the standard treatment. The learned model in G2P2C may be used to improve explainability by providing the user with a visualisation of the forward planning carried out by the algorithm. This may provide insights towards the suggested insulin actions. However, further research and collaboration with clinicians are required to improve the explainability of RL-based APS.

8.3.2 Technology Transfer and Research Translation

RL algorithms require a large amount of data for training. Hence, as an initial step, T1D simulators are required to train the algorithms. The simulators present people with varying glucose dynamics to represent the T1D patient population. Hence, prior to the application on *in-vivo*, *in-silico* training based on a digital twin is required.

However, the *transferability* would also require sufficient real-world training, which could be infeasible and extremely dangerous to be conducted on-line. Hence, an off-line approach to use existing patient data to fine tune the learned control strategy is required. In future work, the use of off-line patient data to fine-tune the glucose dynamics modules in G2P2C to learn a personalised glucose dynamics of the target subject can be explored. The learned model can also be used for simulation based training for the RL algorithm. A fundamental limitation to research in the area of RL-based glucose control algorithms is the restrictions present in current T1D simulators. In future work, with additional engineering effort platforms can be created for RL algorithms to use the latest version of the UVA/PADOVA simulator (2018).

8.3.3 Privacy and Security

Privacy of information and patient security are vital in healthcare applications. APS can be considered as a networked information systems where patient data such as glucose measurements and administered insulin information is stored and a control algorithm used to process this data. Hence, attention is required to ensure the security and privacy of communication related to this system where, user data can be compromised and unauthorised access to the data and manipulation could be life threatening due to the safety critical nature of the application.

Encrypting the control algorithm is an approach to enhance privacy and security and has been explored in various applications with linear controllers [Kogiso and Fujita, 2015]. During the thesis, a collaboration study explored homomorphic encryption for APS operating under a PID controller, which achieved encryption without compromise on performance [Weng et al., 2023]. However, an RL-based control algorithm is non-linear and much more complicated for the encryption procedure. Hence, further research is required to evaluate the security and privacy aspects of RL-based APS.

8.3.4 Ethical Considerations

The application of RL to real-world health applications is a relatively new field of study. Hence, future research is required to explore ethical implications associated to responsibility and accountability. The FDA has identified guidelines for AI/ML applications in healthcare which would be valuable for the future development of glucose control methods. However, further explorations need to be conducted specifically in relation to RL algorithms, since it has the ability to continue learning and improving based on the patient.

8.3.5 Clinical Trials

Clinical trials are required to properly evaluate the proposed G2P2C algorithm. The regulations applicable to the target geography needs to be identified and carefully analysed prior to clinical trials. A human-in-the-loop approach is expected to be

valuable in initial clinical trial settings, where low-risk clinical trials with a small duration can be considered and a clinician added to the control loop to accept or reject the suggested insulin dose of the RL algorithm. This would be valuable to ensure the safety of the patient and the clinical feedback could be a considered as expert demonstrations which is useful for fine-tuning of the RL algorithm. This is expected to be an iterative process to improve the algorithm, its safety, and explainability.

8.4 Summary

The development of RL-based APS requires a multi-disciplinary research effort. In this thesis, CAPSML and GluCoEnv were engineered as tools to facilitate collaboration between these different stakeholders. CAPSML aims to improve the communication between clinicians, lived experience experts, and RL researchers, while GluCoEnv focuses on using engineering to assist the research effort. Both the tools (CAPSML: https://capsml.com/, GluCoEnv: https://github.com/chirathyh/GluCoEnv) are released as open-source under the MIT license along with the codebase of G2P2C (https://github.com/chirathyh/G2P2C) to support the development of RL-APS.

Thesis Conclusion

The thesis introduced the research problem of glucose regulation in T1D in Chapter 1 and presented the background material related to T1D, APS, ML, and RL in Chapters 2 and 3. This thesis builds on the previous research efforts in T1D and RL domains (Figure 9.1). The Chapters 4, 5, 6, and 7 presented the research contributions related to RL-APS and MAPS. In Chapter 8, engineering contributions were presented with future directions for APS. This final chapter concludes the thesis by summarising the research/engineering contributions, highlighting limitations, and presenting the broader research impact of this thesis.

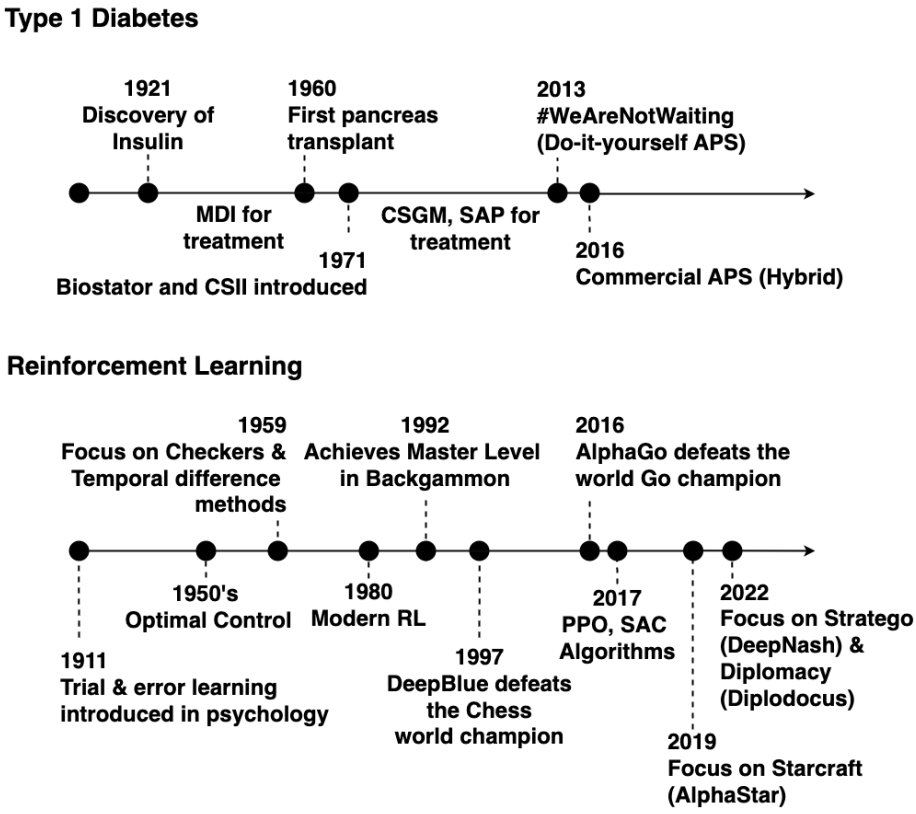


Figure 9.1: A timeline of type 1 diabetes and reinforcement learning milestones.

9.1 Thesis Summary and Contributions

T1D is a chronic disease affecting millions of people worldwide, requiring exogenous insulin administration to maintain glucose homeostasis [DiMeglio et al., 2018]. APS can be identified as the latest treatment strategy, which is an HCL system [Cobelli et al., 2011]. However, they require manual inputs from the user (CHO estimation, meal announcement, and insulin dose calculations), which adds a significant cognitive burden on people with T1D and their carers, while human error affects the performance leading to sub-optimal glucose control (Chapter 2) [Brew-Sam et al., 2021a]. The problem of glucose regulation is challenging due to factors such as complex dynamics, delays, disturbances, and uncertainty [Bothe et al., 2013]. ML and RL techniques have been proven to address these challenges and have also been explored in APS development (Chapter 3) Bothe et al. [2013].

This thesis primarily explored using RL algorithms to design APS to automate the treatment. A novel RL algorithm named G2P2C, capable of eliminating meal announcement and CHO estimation was proposed to reduce the cognitive burden, which outperformed clinical treatment on the adult T1D cohort (Chapter 5). To develop G2P2C, first, an RL problem formulation was proposed, and next, the state-of-the-art RL algorithms were explored to understand their algorithmic characteristics and suitability for glucose regulation (Chapter 4). The RL problem formulation focused on challenging characteristics of real-world problems such as partial observability, delays, and reward function design. A novel method of using non-linear action spaces for RL agents was proposed to tackle the complex action space of the problem, which improved the learning performance, efficiency, and convergence properties of RL agents (Chapter 6). As a secondary objective of the thesis, a systematic literature review was conducted to investigate the use of additional physiological signals (e.g., heart rate, galvanic skin response, and lactate) to design MAPS as an extension to APS. Invasive physiological signals, which were minimally explored in previous work, were analysed in this thesis in the context of different exercise types of T1D to explore valuable physiological signals for glucose control (Chapter 7).

The thesis focused on a multidisciplinary research approach, where feedback from people with T1D, health experience experts, clinicians, engineers, and researchers was incorporated to identify thesis objectives and develop solutions. Furthermore, communication between the stakeholders was improved by introducing two software tools, CAPSML and GluCoEnv. The thesis also provides the codebase of this research as open-source under the MIT license for replication and facilitating the future development of RL-based APS (Chapter 8). The key contributions of this work based on the research objectives are summarised below.

(1) Develop fully automated RL-based APS to alleviate the cognitive burden on people with T1D:

Eliminating Meal Announcement & CHO Estimation. The glucoregulatory system is challenged by the uncertainty of daily meal events and their disturbances to glucose control. Furthermore, the delays in glucose sensing (roughly 6–12 minutes) and insulin action (roughly 5–15 minutes, depending on the type of insulin) add further challenges [Villena Gonzales et al., 2019; Vliebergh et al., 2021]. As a result, existing treatment requires the use of meal announcements typically 20 minutes in advance to reduce uncertainty and the estimation of the CHO content of the meal to quantify the potential disturbance to glucose. However, this process also adds a significant burden on people with T1D [Brew-Sam et al., 2021a]. Hence, this thesis explored the use of RL algorithms, which are suited for stochastic control tasks, and designed a glucose control algorithm, called G2P2C, which does not require such manual inputs.

RL Problem Formulation & Non-linear Action Spaces. In contrast to popular board game environments, applying RL to real-world problems presents many challenges. Hence, this thesis proposed a POMDP-based problem formulation where state/action space and reward function designs were presented. The action space was explored in detail to address the challenges of the large continuous action space and the possible existence of redundant actions due to the specific characteristics of the problem. A novel idea was introduced, where translation functions were proposed to map the action space of the RL agent to the control space of the actuator. This proposed translation was non-linear, in contrast to a more traditional approach of mapping the action and control spaces linearly in RL tasks. It resulted in an improved performance and efficiency of learning while portraying smooth convergence properties.

G2P2C Algorithm. A comparison study of the state-of-the-art RL algorithms for glucose control identified PPO as a suitable candidate algorithm for APS development due to its constrained policy optimisation procedure. However, due to the practical limitations of the RL objective based on an infinite-horizon control problem, catastrophic failures were identified in the PPO algorithm. Hence, a novel algorithm named G2P2C was developed. The G2P2C algorithm was designed based on the state-of-the-art PPO algorithm, where two novel optimisation phases — model learning and planning — were introduced. The model learning phase integrated into G2P2C learned a glucose dynamics model, which was used in the planning phase to plan for the short horizon and fine-tune the policy. This mechanism reduced catastrophic failures and also improved performance. As a result, G2P2C outperformed the basal-bolus clinical treatment for the adult T1D cohort, even though no meal announcements and CHO estimations were provided.

(2) Explore the use of additional physiological variables to develop MAPS:

MAPS. The widespread nature of wearable devices and increased attention towards developing invasive sensors provide an opportunity to use valuable additional information for glucose control. Hence, in this thesis, a systematic literature review was conducted to identify additional physiological signals that could improve glucose control. The review identified that invasive physiological signals were explored less, while additional physiological signals were mainly used to address glucose control challenges associated with exercise. Similarly to meals, exercise is a daily activity that affects glucose control, and hence, its accommodation in existing treatments was challenging due to the rapidly changing insulin needs related to different types of exercise. Hence, invasive physiological signals were analysed in the context of different types of exercise (MIE, HIE, and RE). An MTE approach was used to model the physiological signals to learn the inter-relations of the glucoregulatory system. The analysis validated lactate as an important biomarker in estimating exercise events, while noradrenaline was observed to be valuable in differentiating RE from HIE and MIE. This work is expected to be valuable towards the development of MAPS, which is considered the next generation of APS.

(3) Develop software applications/tools to support T1D education and research:

CAPSML & GluCoEnv. Development of RL based-APS requires the combined effort of people with T1D, clinicians, health experience experts, computer science researchers, and engineers. However, in existing research, the lack of communication between the groups leads to solutions that are practically infeasible to implement and are also evaluated on non-realistic meal protocols. A significant barrier also exists in communicating the concepts of RL to people from non-RL backgrounds. Hence, the demonstration tool CAPSML was designed to improve the communication between technical and non-technical audiences. The feedback from this tool is expected to be valuable in future RL-based APS development. Furthermore, GluCoEnv was engineered as a simulation tool to enhance the experimentation process in RL-based APS. The introduced tools aim to address human and technical barriers associated with developing RL-based APS.

Comparison with Prior Work. In contrast to most previous work on RL-based APS [Sun et al., 2018; Myhre et al., 2018; Ngo et al., 2018a,b; Sun et al., 2019], this thesis aimed to develop fully automated RL-based APS, which eliminates the requirement of meal announcement and CHO estimation. Furthermore, this thesis identified limitations in the handful of previous automated RL-based APS, where the development focus has been on low CHO meals [Fox and Wiens, 2019; Fox et al., 2020], evaluated on adult-only T1D cohorts [Lee et al., 2020b], and resulted in over-personalised solutions [Lee et al., 2020b]. Hence, this thesis focused on a generalised RL-based APS solution evaluated on multiple T1D cohorts and designed for realistic high CHO meal scenarios. In addition, a much stricter definition of catastrophic failures was

used in this study. The novel G2P2C algorithm proposed in this thesis outperformed the SAC-based hybrid approach proposed in Lim et al. [2021] for adult and adolescent cohorts. However, the performance was lower in the adult cohort compared to Lee et al. [2020b]. The comparison is limited due to the different experimental settings used, highlighting the need for appropriate APS development benchmarks. The replication of previous studies was also impossible due to the unavailability of source code and associated parameters. Hence, following best practice, basal-bolus clinical treatment strategies were implemented, and standard clinical metrics were used for the evaluation, where G2P2C outperformed an ideal basal-bolus strategy in the adult cohort. However, G2P2C can be further improved to reduce failures and future work required to explore the explainability and transferability aspects of the algorithm. To facilitate the future development of RL-based APS and replication, the codebase of this thesis is open-source under the MIT license.

MAPS presents a promising direction for future research on glucose regulation. In previous studies Kudva et al. [2014] analysed the clinical importance of additional physiological signals, while Cinar [2017] and Patek [2019] analysed the current limitations of APS to approach MAPS. In contrast, the systematic literature review conducted in this thesis focused on the technical aspects of MAPS development, where types of additional physiological signals used, control techniques, architectures, and validation methodologies were explored. Furthermore, the MTE-based analysis on invasive physiological signals in the context of different exercise types presented in this thesis augments previous work by Paldus et al. [2022b] on a relatively less explored domain.

9.2 Limitations

T1D Simulator. The RL-based APS proposed in this research used a T1D simulator based on the FDA-approved UVA/PADOVA 2008 model [Kovatchev et al., 2009]. The best practice in developing glucose control algorithms is using simulators for pre-clinical proof-of-concept testing. However, these simulators have limitations in replicating real-world scenarios (e.g., simulating rescue CHO used in the real-life, impact of stress, exercise). Hence, this research was restricted to an *in-silico* analysis and future clinical trials are required to assess the performance of the proposed algorithms. Furthermore, additional physiological signals were identified and analysed as part of this research. However, these additional inputs were not integrated into the control algorithm due to the limitations of the simulators to model the effect of these physiological signals. Furthermore, using existing simulators for developing RL-based APS presents numerous engineering challenges (e.g., integration of RL algorithms and running multiple agents in parallel). Hence, in the thesis, a GPU-based simulator was designed based on the UVA/PADOVA 2008 model to facilitate the development of RL-based APS.

Evaluation Benchmarks. The field of glucose control algorithms lacks appropriate benchmarks for comparison between studies. This is mainly due to the various experimental settings (simulators, meal protocols, T1D cohorts) used. Hence, research on glucose control often benchmarks the performance of algorithms against standard clinical treatment approaches and guidelines, as presented in this work. In this thesis, basal-bolus clinical treatment strategies were replicated as the gold standard for evaluation due to their widespread recognition among clinicians. This benchmark, however, was challenging as they used meal announcement and CHO estimation, whereas the RL-based APS proposed in this work eliminated this requirement to reduce the cognitive burden on people with T1D. Furthermore, it is essential to highlight that replicating previous RL-based APS solutions was impossible due to the unavailability of source code and hyperparameters. The replication of RL-based studies also requires significant computing resources to be trained from scratch. Hence, the codebase of this study was released publicly for the reproducibility of the work, along with trained models released as part of an online demonstration tool under the MIT license to facilitate future work in this field.

Exploring State-Of-The-Art RL Algorithms. In recent years and especially during the period of this thesis, there has been an explosion of RL algorithms being designed. The training and fine-tuning hyperparameters for RL algorithms targeting the 20-*in-silico* subjects present in this thesis requires significant computational resources. Hence, in this thesis, the most widely accepted and explored the state-of-the-art RL algorithms A2C (2016) [Mnih et al., 2016], PPO (2017) [Schulman et al., 2017], and SAC (2018) [Haarnoja et al., 2018] were explored. Variations of the above and more recent algorithms such as CPO (Constrained Policy Optimisation, 2017) [Achiam et al., 2017], AlphaZero (2017) [Silver et al., 2017], ME-TRPO (Model-Ensemble Trust Region Policy Optimisation, 2018) [Kurutach et al., 2018], MBPO (Model-Based Policy Optimisation, 2019) [Janner et al., 2019], GAMPS (Gradient-Aware Model-based Policy Search, 2020) [D'Oro et al., 2020], and MuZero (2020) [Schrittwieser et al., 2020] were not explored in this thesis. However, the ideas present in these latest algorithms would be beneficial to improve G2P2C in future work.

G2P2C Algorithm. The successful real-world application of the proposed G2P2C algorithm requires further research on the areas of safety, personalisation, transferability of the *in-silico* learned strategy to real-life, and explainability of the control strategy. The design of G2P2C considered all these aspects where the learned glucose dynamics model could be used for developing safety modules (e.g., glucose prediction, hypoglycaemia alarm systems) and personalising learnt glucose control strategies to real individuals by using off-line data to fine-tune the glucose dynamics model. The demonstration tool CAPSML was introduced as a first step to improve the explainability and communication between researchers, clinicians, and people with T1D. This study did not analyse the impact of the model error of the learned glucose dynamics model on the control performance. Future work could explore the effect of using glucose prediction for different time horizons (e.g., 30, 60 min-

utes) in contrast to the one-step (5-minute) ahead predictions used in this study. The planning mechanism of G2P2C can be used to further improve the explainability by presenting an analysis of different glucose control scenarios. This research is reserved for future work.

Action Space Analysis. This thesis identified the state-of-the-art PPO algorithm as suitable for the glucose regulation application, which requires continuous control and steady learning. Hence, the non-linear action space analysis was only conducted based on PPO to reduce the compute resource requirements. However, it is expected that the identified benefits are also applicable to other similar on-policy RL algorithms, such as A2C. The results draw a promising line of research to explore the contribution of non-linear action spaces on other on-policy and off-policy RL algorithms and also on other RL problems. The use of a linear action space does not impose any bias in the learning process, as all actions are equally probable. In contrast, the proposed action space formulation used prior knowledge about the insulin action distribution to facilitate better learning which incorporates the bias of the current clinical practice in T1D insulin treatment. This introduced bias was demonstrated as valuable for the glucose regulation task to achieve effective and efficient control. However, the effect of the bias in the proposed approach could be detrimental in problem domains with limited expert knowledge, and future research could focus on methods to identify the most suitable translation functions to design the target non-linear action spaces.

MAPS Survey and Analysis. The systematic literature review conducted as part of this thesis analysed existing APS designs to identify the types of input variables used, control techniques, architectures, and validation methodologies. Its scope was limited to studies that proposed APS. However, research studies exist that aim to identify relationships between various physiological signals and T1D. Surveying and identifying such relationships would benefit the development of MAPS. The study on analysing invasive physiological signals during exercise was limited to 30 participants on a closed-loop treatment strategy, while the biochemical data were sampled every 20 minutes. These limitations affect the identified relationships and the accuracy of the calculated information flow delays. However, even though larger datasets with better resolution are valuable, such datasets are currently unavailable.

9.3 Broader Impact Statement

This PhD thesis was conducted as a multidisciplinary research project with multiprofessional collaboration and collaborative co-creation with people with T1D, clinicians, health experience experts, engineers, and computer science researchers. The collaboration among these stakeholders enabled this thesis to identify the importance of exploring methods to reduce the cognitive burden and the potential of additional

physiological signals. In addition, the co-creation activities with clinicians helped fine-tune reward functions, incorporating real-life problem constraints in designing the G2P2C algorithm. The societal impact of this research is broadly towards diabetes and ML communities.

Diabetes Community. The control algorithms developed in this research aim to improve current treatment by eliminating meal announcements and CHO estimation to reduce the cognitive burden on people with T1D. In addition, the proposed G2P2C algorithm and the CAPSML tool allow clinicians to understand and learn about the RL-based APS and their potential to improve future treatment strategies while being actively involved in the development process.

ML Community. Most existing research on RL focuses on board games and computer games. In contrast, this thesis explored using RL algorithms towards the real-world glucose control problem. Real-world RL systems are presented with many challenges, which are also present in the problem of glucose regulation. Hence, glucose control can be identified as a highly suitable application for developing real-world RL algorithms. This thesis explored solutions to these challenges and proposed a novel modular RL algorithm, G2P2C, which showed promise to be used in real-world applications and thus highlights the importance of developing such algorithms to suit real-world problems. Hence, the work presented in this thesis is valuable towards applying RL methods to real-world applications.

Popular RL application domains have extensive online resources for researchers and developers, where the problem is clearly formulated and open-source codebases released with hyperparameter settings and documentation. However, applying RL to glucose control presents many engineering challenges, and minimal associated resources are available. Hence, the codebase of this thesis was open-sourced, and a GPU-based T1D simulator was designed to address engineering constraints. These efforts reduce the barriers to entry and motivate RL researchers to explore this real-world problem. Hence, these contributions are valuable for the ML community.

Ethical and Legal Implications. RL-based APS is a relatively new area of research. Hence, future work and collaborations are required. Open-source contributions are valuable towards the progress of the field.

"When you put together open medicine, open science, open access, open source, and open data—Open^5—all sorts of new channels of research activity become available, and existing ones become exponentially more powerful."
-Eric Topol, The Patient Will See You Now: The Future of Medicine is in Your Hands [Topol, 2015, p.211].

However, as identified in Chapter 8, open-source initiatives such as #WeAreNotWaiting have commenced developing do-it-yourself APS. These community-driven

projects use open-source resources and crowd-sourced data to develop control algorithms. Hence, releasing open-source resources presents a risk to people with T1D, as they can be used to design control algorithms. Due to these ethical implications, the resources related to the trained models (e.g., model files) were not released to the public in this research. Instead, to encourage the co-creation process, the designed online demonstration tool CAPSML (MIT license) allows the public to test the trained algorithms through a simulation interface. However, the codebase is open-source under the MIT license, allowing researchers to train their own algorithms and are encouraged to collaborate with the project for additional information and resources.

Proper ethics approvals were obtained for the work conducted during this thesis (Australian National University, Human Research Ethics Committee, Protocol Number: 2020/411 and Protocol Number: 2020/534, St. Vincent's Hospital, Melbourne, Human Research Ethics Committee Reference Number HREC 021/18). In parallel with the research and development process, work on ethical and legal guidelines is a must to realise RL-based APS. In addition, clarity in the regulatory framework of using RL for healthcare and conducting clinical trials would benefit the RL-APS development process.

Research Outreach Initiatives. In addition to the research publications targeting the academic community in T1D and ML, several research outreach initiatives were conducted to communicate the concepts of RL and APS to the general public. The research outcomes of the thesis was presented at the World Diabetes Day - 2022 event in Canberra and will be demonstrated at the Medical Informatics in Europe Conference 2023 and during the Vitalis 2023 eHealth exhibition on May 2023.

9.4 Final Remarks

The importance of a multidisciplinary research effort was highlighted throughout the work of this thesis. The guidance and contribution of people with T1D, clinicians, and health experience experts were valuable to identify the cognitive burden in existing treatment as an important area of improvement. The thesis was limited to an *in-silico* analysis. However, the proposed G2P2C algorithm shows promise as a candidate algorithm for glucose control. The tools developed during this thesis is expected to be valuable for future RL-based APS development. The thesis also identified important and interesting areas of research in RL, which would be valuable in future applications and real-world RL systems.

Appendix A

Sample Glucose Control Simulations for G2P2C.

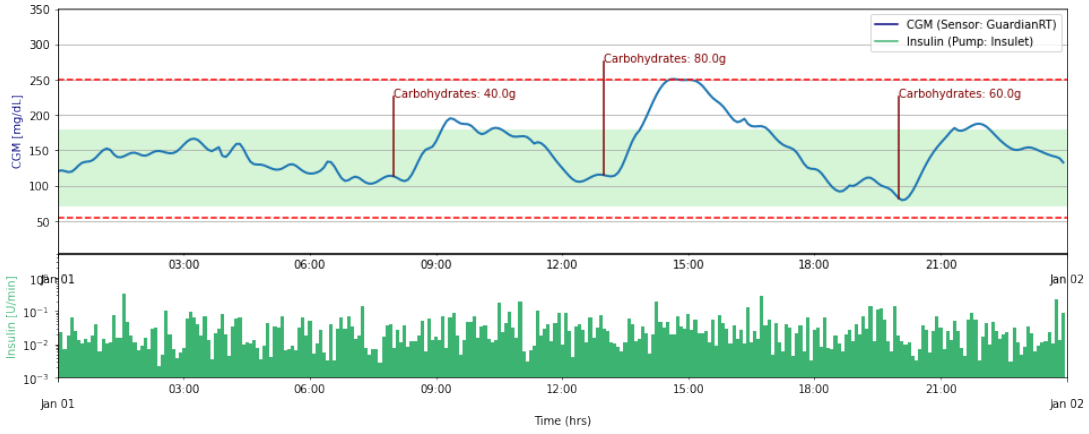


Figure A.1: Sample glucose simulation for Adult0 using G2P2C.

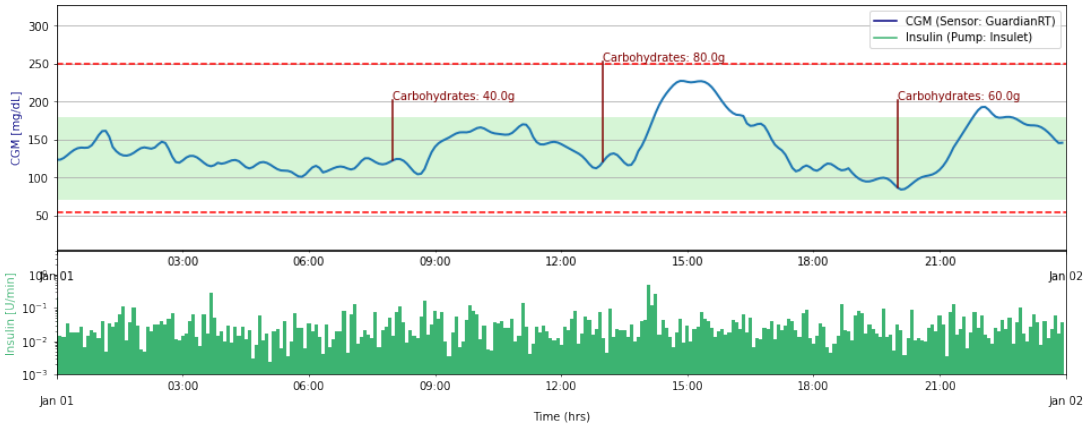


Figure A.2: Sample glucose simulation for Adult1 using G2P2C.

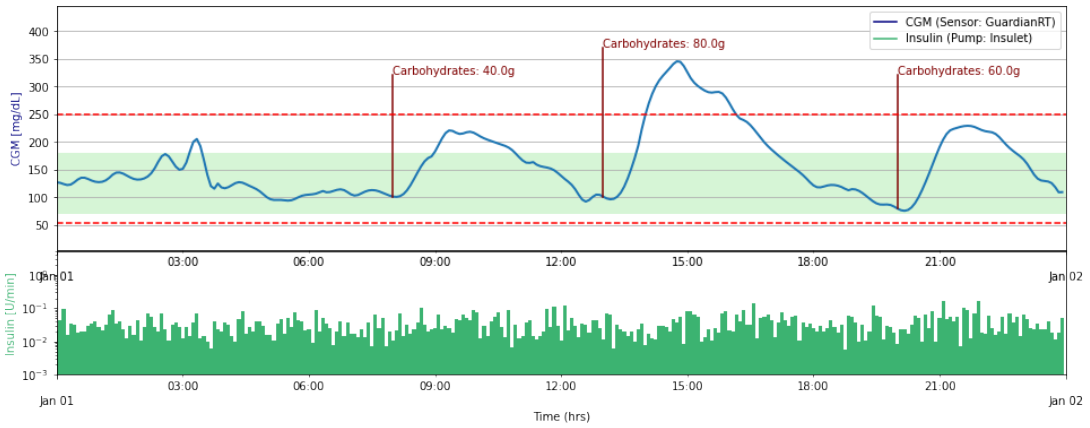


Figure A.3: Sample glucose simulation for Adult2 using G2P2C.

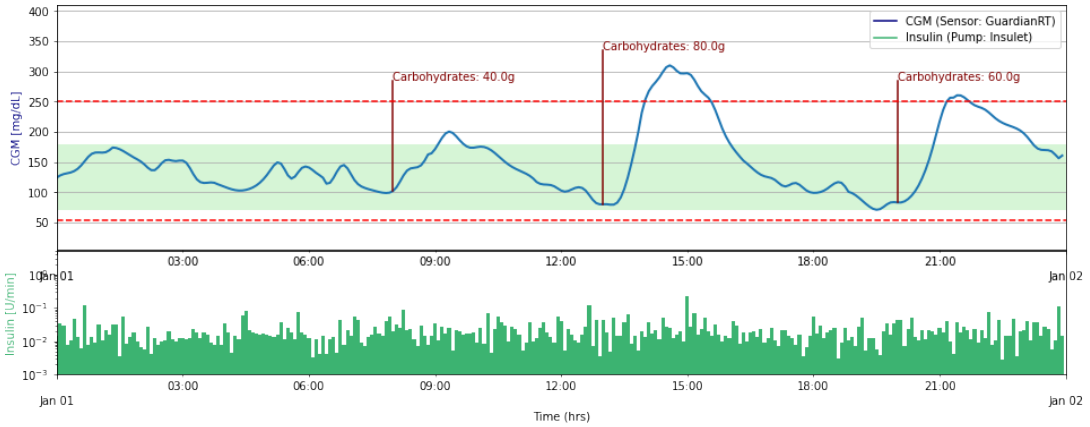


Figure A.4: Sample glucose simulation for Adult3 using G2P2C.

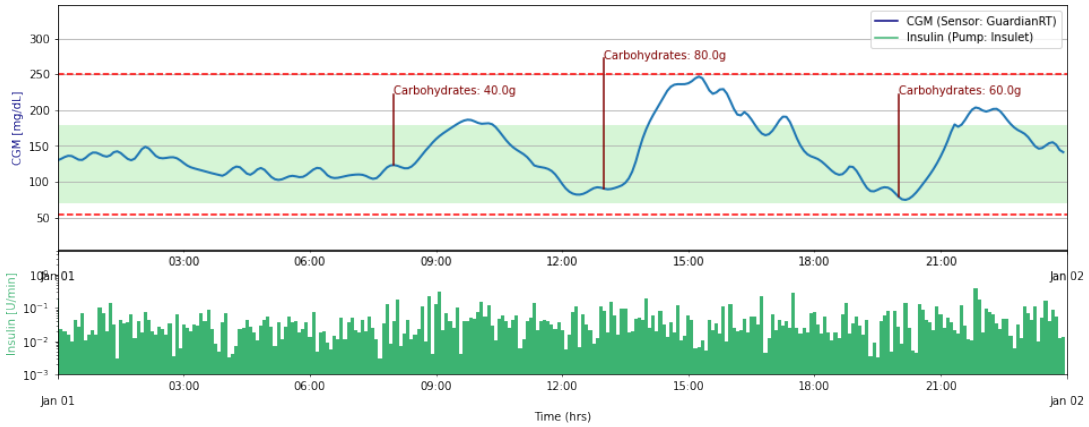


Figure A.5: Sample glucose simulation for Adult4 using G2P2C.

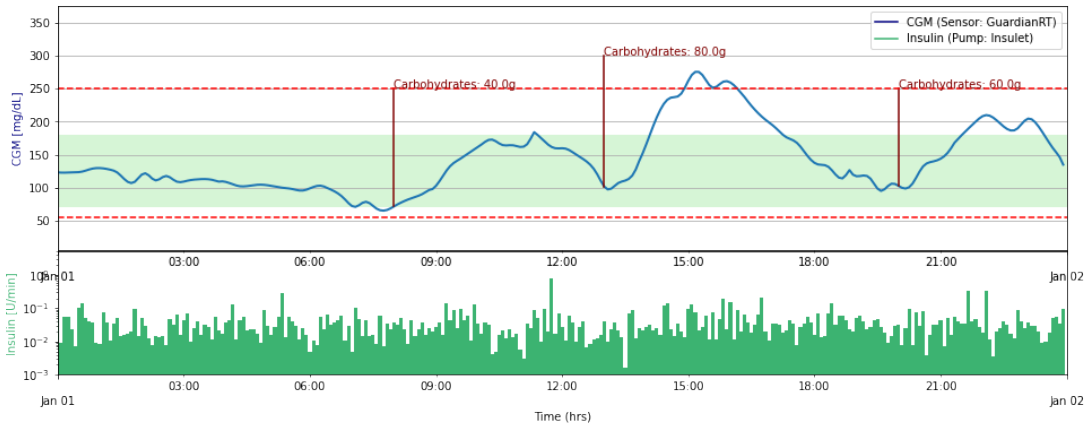


Figure A.6: Sample glucose simulation for Adult5 using G2P2C.

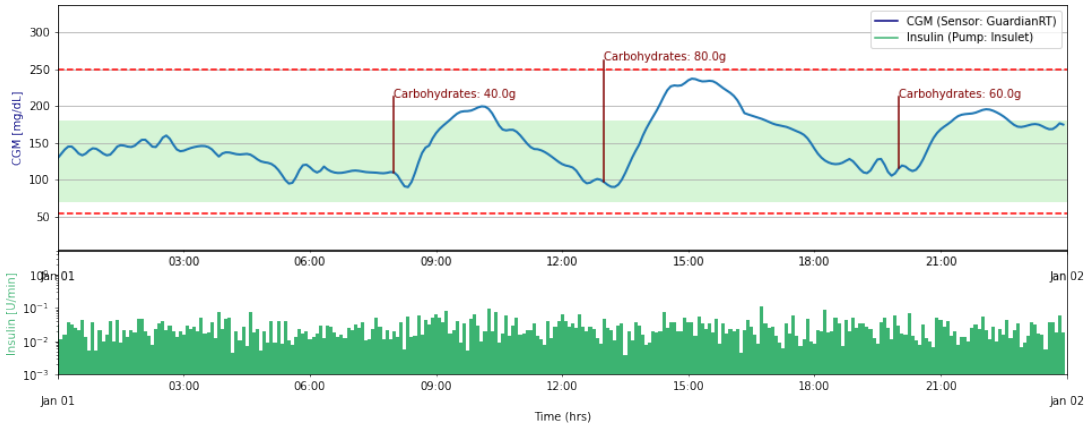


Figure A.7: Sample glucose simulation for Adult6 using G2P2C.

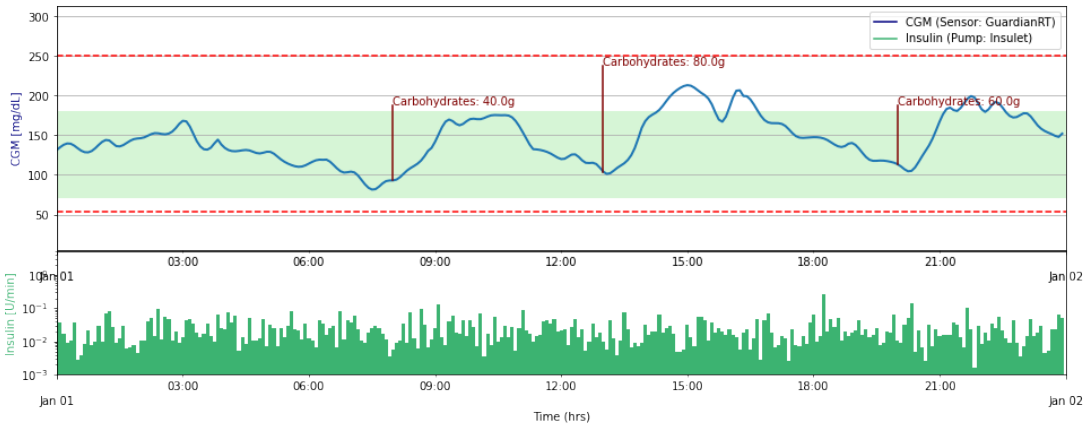


Figure A.8: Sample glucose simulation for Adult7 using G2P2C.

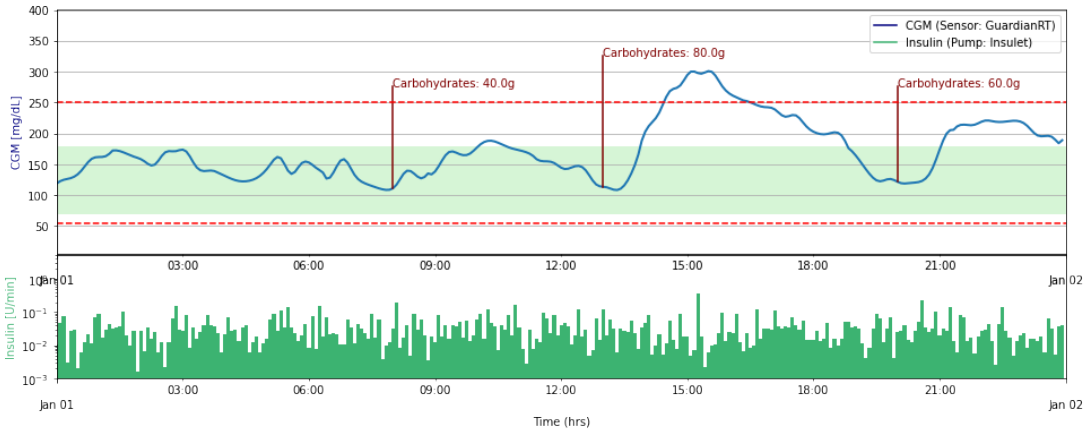


Figure A.9: Sample glucose simulation for Adult8 using G2P2C.

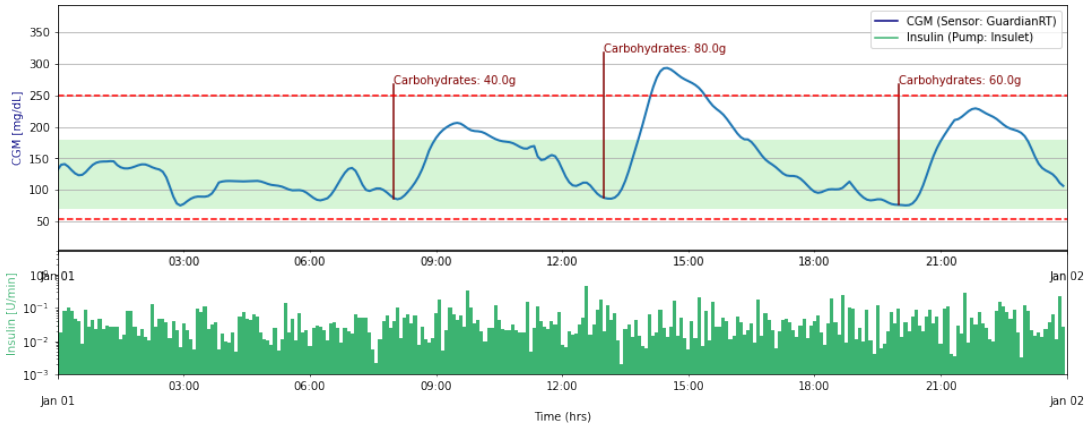


Figure A.10: Sample glucose simulation for Adult9 using G2P2C.

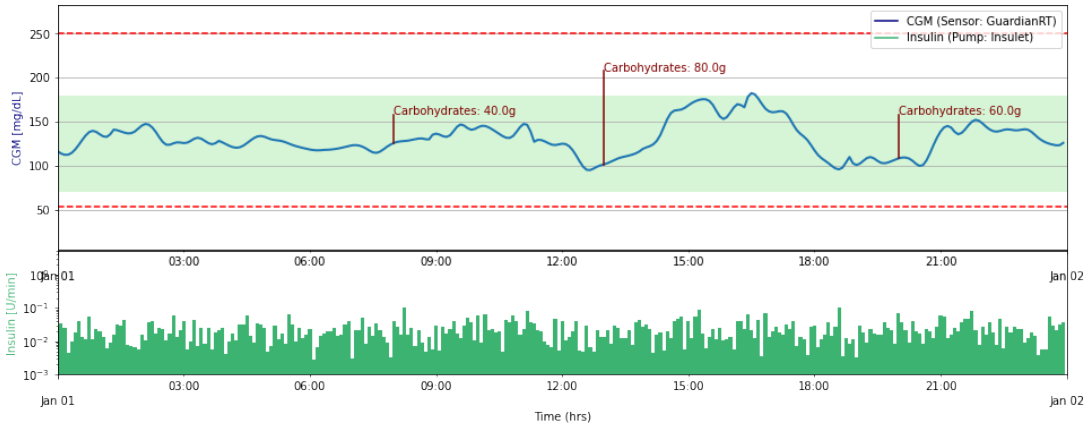


Figure A.11: Sample glucose simulation for Adolescent0 using G2P2C.

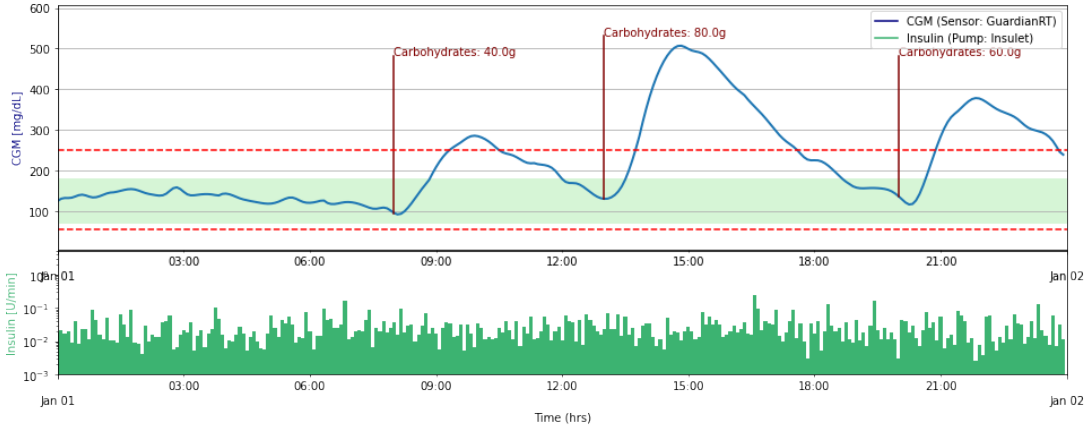


Figure A.12: Sample glucose simulation for Adolescent1 using G2P2C.

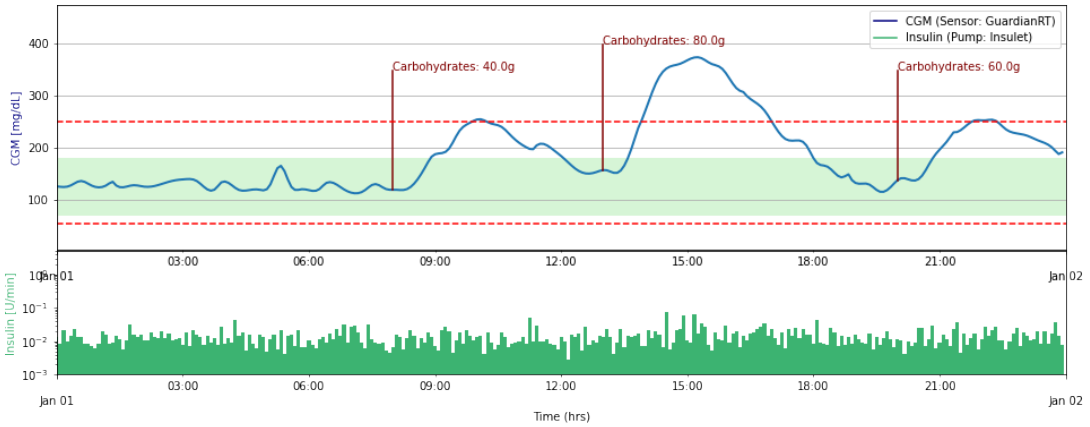


Figure A.13: Sample glucose simulation for Adolescent2 using G2P2C.

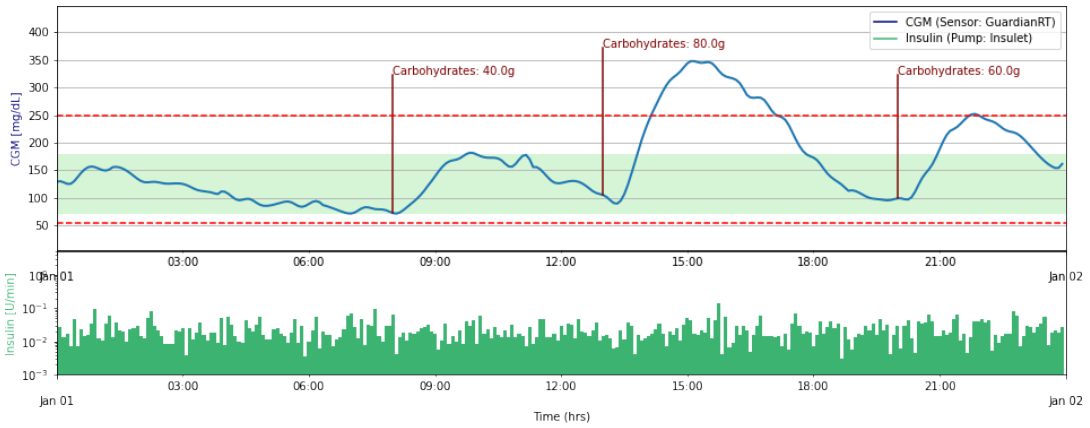


Figure A.14: Sample glucose simulation for Adolescent3 using G2P2C.

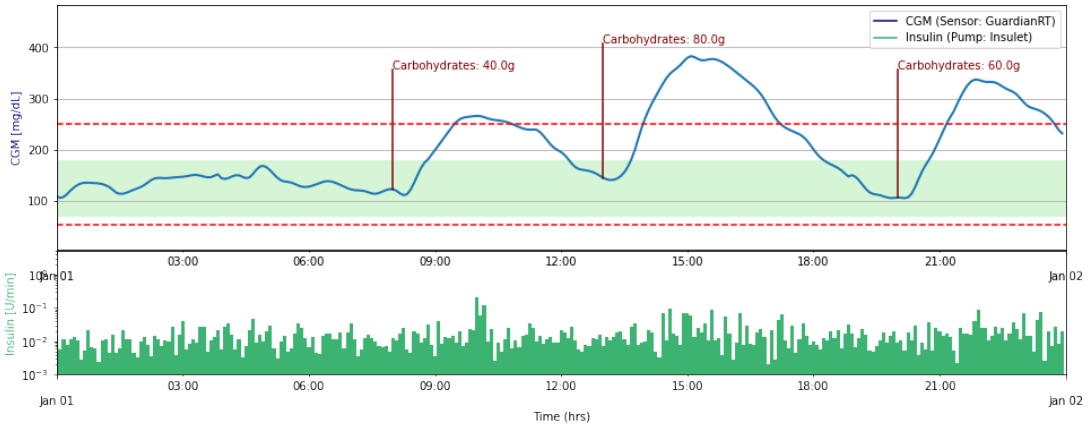


Figure A.15: Sample glucose simulation for Adolescent4 using G2P2C.

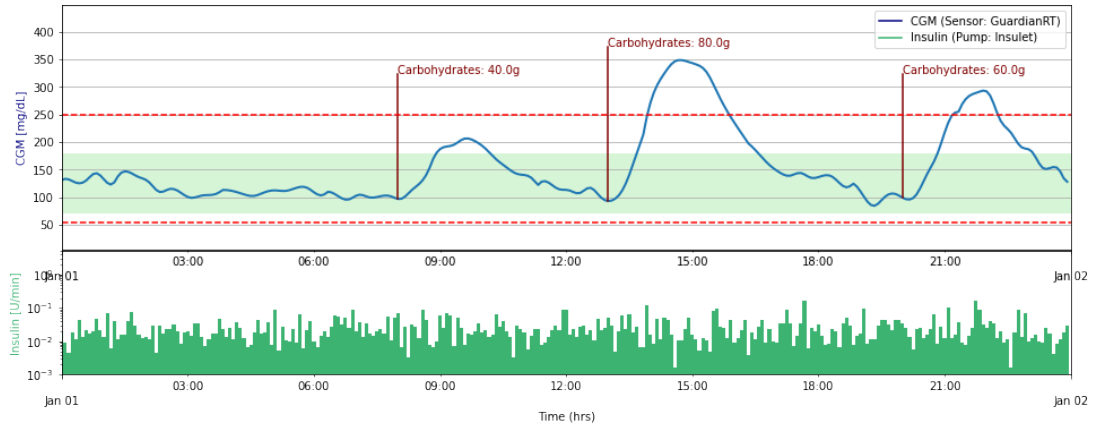


Figure A.16: Sample glucose simulation for Adolescent5 using G2P2C.

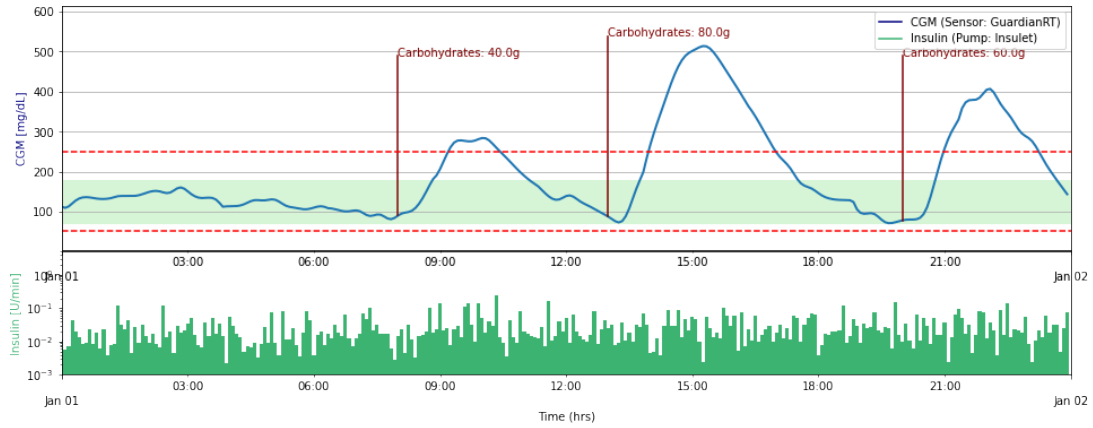


Figure A.17: Sample glucose simulation for Adolescent6 using G2P2C.

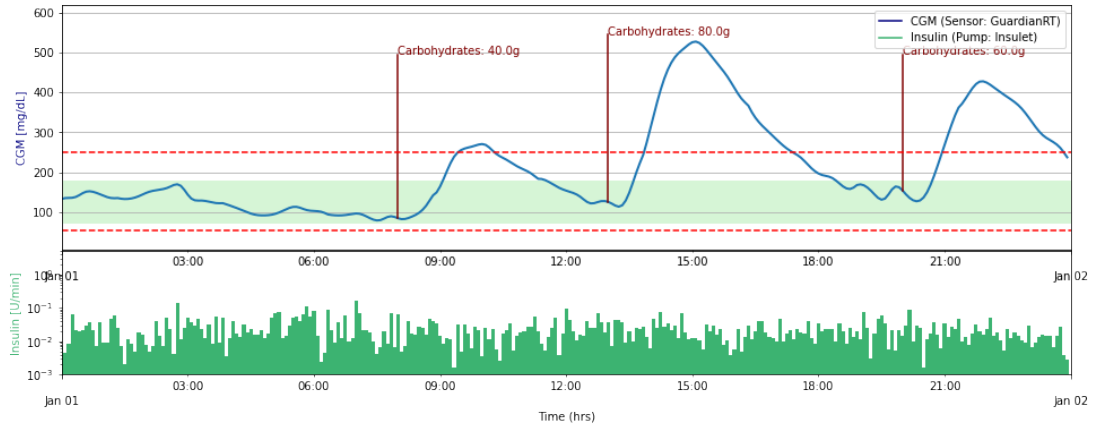


Figure A.18: Sample glucose simulation for Adolescent7 using G2P2C.

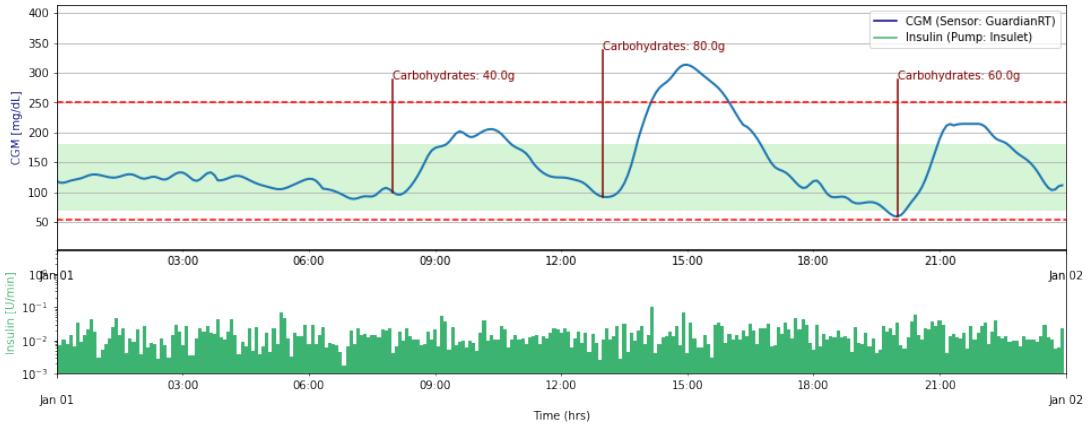


Figure A.19: Sample glucose simulation for Adolescent8 using G2P2C.

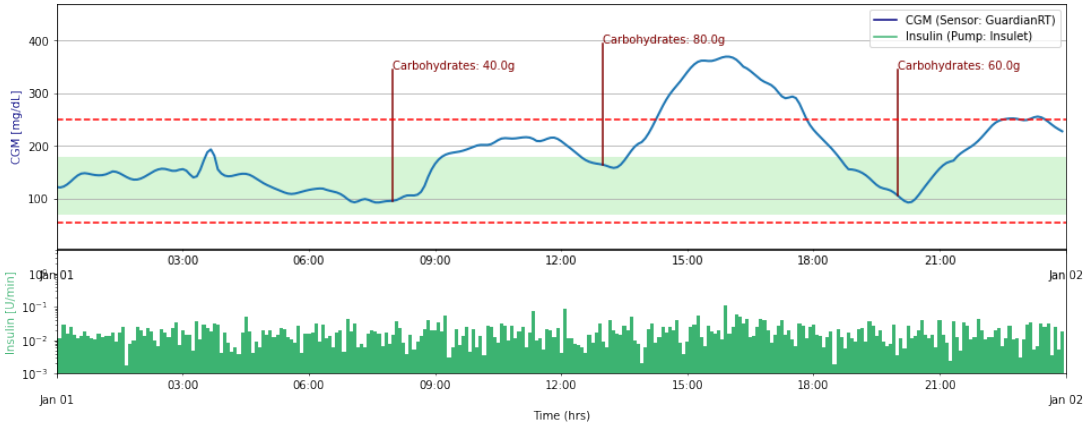


Figure A.20: Sample glucose simulation for Adolescent9 using G2P2C.

Appendix B

Systematic Literature Review Resources

B.1 Search Query Formulation

Different strategies were explored to develop a suitable search query with the help of a research librarian. It was identified that the studies were not able to be simply queried through the use of the Medical Subject Headings (MeSH) terms due to the following constraints:

- All papers had not been indexed using MeSH terms (e.g., diabetes mellitus, type 1 and pancreas, artificial).
- Some MeSH terms (e.g., Devices, Wearables) had been introduced recently whereas this survey focused on literature for the period 2005–2020.

Hence, the proposed query (Table 7.1) was designed to encompass all studies related to APS and control. The queries were not restricted by terms, such as "wearables, physiological signals", in order to ensure that none of the potential studies was omitted. Hence, instead all the 1388 results from the general queries were manually screened to identify the appropriate studies.

B.2 CASP Quality Assessment of Selected Studies

CASP (Critical Appraisal Skills Programme) Questions:

- Question 1: Was there a clear statement of the aims of the research?
- Question 2: Is a qualitative methodology appropriate?
- Question 3: Was the research design appropriate to address the aims of the research?
- Question 4: Was the recruitment strategy appropriate to the aims of the research?
- Question 5: Was the data collected in a way that addressed the research issue?

- Question 6: Has the relationship between researcher and participants been adequately considered?
- Question 7: Have ethical issues been taken into consideration?
- Question 8: Was the data analysis sufficiently rigorous?
- Question 9: Is there a clear statement of findings?
- Question 10: How valuable is the research?

Table B.1: CASP quality assessment.

Study	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Main Studies										
Quiroz and Femat [2010]	Y	Y	Y	NA	NA	CT	NA	Y	Y	Y
Quiroz et al. [2011]	Y	Y	Y	NA	NA	CT	NA	Y	Y	Y
Khan et al. [2013]	Y	Y	Y	NA	NA	CT	NA	Y	Y	Y
Qaisar et al. [2012]	Y	Y	Y	NA	NA	CT	NA	Y	Y	Y
Stenerson et al. [2014]	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
DeBoer et al. [2017]	Y	Y	Y	Y	Y	CT	Y	Y	Y	Y
Jacobs et al. [2015]	Y	Y	Y	Y	Y	CT	CT	Y	Y	Y
Resalat et al. [2019b]	Y	Y	Y	Y	Y	CT	CT	Y	Y	Y
Turksoy et al. [2013a]	Y	Y	Y	Y	Y	CT	N	Y	Y	Y
Turksoy et al. [2013d]	Y	Y	Y	Y	Y	CT	CT	Y	Y	Y
Hajizadeh et al. [2019]	Y	Y	Y	NA	NA	CT	NA	Y	Y	Y
Clinical Studies										
Breton et al. [2014]	Y	Y	Y	Y	Y	CT	Y	Y	Y	Y
Jacobs et al. [2016]	Y	Y	Y	Y	Y	CT	Y	Y	Y	Y
Castle et al. [2018]	Y	Y	Y	Y	Y	CT	Y	Y	Y	Y
Turksoy et al. [2013b]	Y	Y	Y	Y	Y	CT	CT	Y	Y	Y
Turksoy et al. [2014]	Y	Y	Y	Y	Y	CT	Y	Y	Y	Y
Turksoy et al. [2018]	Y	Y	Y	Y	Y	CT	Y	Y	Y	Y

Y = YES, N = NO, CT = CAN'T TELL, NA = NOT APPLICABLE.

B.3 Patents Associated with MAPS

Table B.2: Patents associated with MAPS.

Patent ID	**Patent Name**	**Application Granted**	**Inventor**
US8690820B2 [Cinar and Oruklu, 2014]	Automatic insulin pumps using recursive multivariable models and adaptive control algorithms.	2014-04-08	Inventors: Ali Cinar, Meriyan Oruklu. Current Assignee: Illinois Inst of Technology.
US10646650B2 [Cinar et al., 2020]	Multivariable artificial pancreas method and system.	2020-05-12	Inventors: Ali Cinar, Kamuran Turksoy, Iman Hajizadeh. Current Assignee: Illinois Inst of Technology.

Bibliography

Abbeel, P. and Ng, A. Y., Apprenticeship Learning via Inverse Reinforcement Learning, *21st International Conference on Machine Learning*, Proceedings (pp. 1), ACM, 2004. doi:https://doi.org/10.1145/1015330.1015430. 31

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P., Constrained Policy Optimisation, *34th International Conference on Machine Learning*, Proceedings (pp 70:22–31), PMLR, 2017. Available: https://proceedings.mlr.press/v70/achiam17a.html. 31, 128

Akkaya, I.; Andrychowicz, M.; Chociej, M.; Litwin, M.; McGrew, B.; Petron, A.; Paino, A.; Plappert, M.; Powell, G.; Ribas, R.; et al., Solving Rubik's Cube with a Robot Hand, *arXiv:1910.07113*, 2019. doi:https://doi.org/10.48550/arXiv.1910.07113. 30, 32

Ananda Basu, Relative Efficacy of Single-, Bi-, and Tri-Hormonal Closed-Loop Control Systems, Endocrinology, Diabetes, Metabolism, and Nutrition, Mayo Clinic, [Online], 2020. Available: https://mayoclinic.pure.elsevier.com/en/projects/relative-efficacy-of-single-bi-and-tri-hormonal-closed-loop-contr. [Accessed on 09/16/2020]. 21

Andrychowicz, M.; Raichuk, A.; Stańczyk, P.; Orsini, M.; Girgin, S.; Marinier, R.; Hussenot, L.; Geist, M.; Pietquin, O.; Michalski, M.; Gelly, S.; and Bachem, O., What matters for on-policy deep actor-critic methods? a large-scale study, *International Conference on Learning Representations*, 2021. Available: https://openreview.net/forum?id=nIAxjsniDzg. 32

Anthimopoulos, M.; Dehais, J.; Shevchik, S.; Ransford, B. H.; Duke, D.; Diem, P.; and Mougiakakou, S., Computer Vision-based Carbohydrate Estimation for Type 1 Patients with Diabetes Using Smartphones, *Journal of Diabetes Science and Technology*, 9, 3, 507–515, SAGE Publications, 2015. doi:10.1177/1932296815580159. 22

Anthony, T.; Tian, Z.; and Barber, D., Thinking Fast and Slow with Deep Learning and Tree Search, *Advances in Neural Information Processing Systems*, 30, 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/d8e1344e27a5b08cdfd5d027d9b8d6de-Paper.pdf. 33

Arjona-Medina, J. A.; Gillhofer, M.; Widrich, M.; Unterthiner, T.; Brandstetter, J.; and Hochreiter, S., Rudder: Return Decomposition for Delayed Rewards, *Advances in Neural Information Processing Systems*, 32,

2019. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/16105fb9cc614fc29e1bda00dab60d41-Paper.pdf. 31

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al., Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI, *Information Fusion*, 58, 82–115, Elsevier, 2020. Available: https://doi.org/10.1016/j.inffus.2019.12.012. 120

Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A., Deep Reinforcement Learning: A Brief Survey, *IEEE Signal Processing Magazine*, 34, 6, 26–38, IEEE, 2017. doi:10.1109/MSP.2017.2743240. 29

Atlas, E.; Nimri, R.; Miller, S.; Grunberg, E. A.; and Phillip, M., MD-logic Artificial Pancreas System: A Pilot Study in Adults with Type 1 Diabetes, *Diabetes Care*, 33, 5, 1072–1076, American Diabetes Association, 2010. doi:10.2337/dc09-1830. 19

Bakhtiani, P. A.; El Youssef, J.; Duell, A. K.; Branigan, D. L.; Jacobs, P. G.; Lasarev, M. R.; Castle, J. R.; and Ward, W. K., Factors Affecting the Success of Glucagon Delivered During an Automated Closed-Loop System in Type 1 Diabetes, *Journal of Diabetes and its Complications*, 29, 1, 93–98, Elsevier, 2015. doi:10.1016/j.jdiacomp.2014.09.001. 103, 105

Bakhtin, A.; Wu, D. J.; Lerer, A.; Gray, J.; Jacob, A. P.; Farina, G.; Miller, A. H.; and Brown, N., Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning, *arXiv:2210.05492*, 2022. doi:https://doi.org/10.48550/arXiv.2210.05492. 30

Bansal, S.; Calandra, R.; Chua, K.; Levine, S.; and Tomlin, C., MBMF: Model-Based Priors for Model-Free Reinforcement Learning, *arXiv:1709.03153*, 2017. doi:https://doi.org/10.48550/arXiv.1709.03153. 28, 29, 33

Barnard, K. D.; Wysocki, T.; Ully, V.; Mader, J. K.; Pieber, T. R.; Thabit, H.; Tauschmann, M.; Leelarathna, L.; Hartnell, S.; Acerini, C. L.; et al., Closing the Loop in Adults, Children and Adolescents with Suboptimally Controlled Type 1 Diabetes Under Free Living Conditions: A Psychosocial Substudy, *Journal of Diabetes Science and Technology*, 11, 6, 1080–1088, SAGE Publications, 2017. doi:10.1177/1932296817702656. 109

Battelino, T.; Danne, T.; Bergenstal, R. M.; Amiel, S. A.; Beck, R.; Biester, T.; Bosi, E.; Buckingham, B. A.; Cefalu, W. T.; Close, K. L.; et al., Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range, *Diabetes care*, 42, 8, 1593–1603, American Diabetes Association, 2019. doi:10.2337/dci19-0028. 12

Bekiari, E.; Kitsios, K.; Thabit, H.; Tauschmann, M.; Athanasiadou, E.; Karagiannis, T.; Haidich, A. B.; Hovorka, R.; and Tsapas, A., Artificial Pancreas

Treatment for Outpatients with Type 1 Diabetes: Systematic Review and Meta-Analysis, 361, BMJ Publishing Group, 2018. doi:10.1136/bmj.k1310. 17

Bengio, Y.; Goodfellow, I.; and Courville, A., 2017. *Deep Learning*, vol. 1. MIT Press Cambridge, MA, USA. 22, 23, 24

Bequette, B. W., Algorithms for a Closed-Loop Artificial Pancreas: The Case for Model Predictive Control, *Journal of Diabetes Science and Technology*, 7, 6, 1632–1643, Sage Publications, 2013. doi:10.1177/193229681300700624. 2, 17, 18, 37

Bergenstal, R. M.; Tamborlane, W. V.; Ahmann, A.; Buse, J. B.; Dailey, G.; Davis, S. N.; Joyce, C.; Peoples, T.; Perkins, B. A.; Welsh, J. B.; et al., Effectiveness of Sensor-Augmented Insulin-Pump Therapy in Type 1 Diabetes, *New England Journal of Medicine*, 363, 4, 311–320, Massachusetts Medical Society, 2010. doi:10.1056/NEJMoa1002853. 50

Bergman, R. N.; Ider, Y. Z.; Bowden, C. R.; and Cobelli, C., Quantitative Estimation of Insulin Sensitivity, *American Journal of Physiology Endocrinology Metabolism and Gastrointestinal Physiology*, 5, 6, 1979. doi:10.1152/ajpendo.1979.236.6.e667. 103, 105

Bergman, R. N.; Phillips, L. S.; and Cobelli, C., Physiologic Evaluation of Factors Controlling Glucose Tolerance in Man: Measurement of Insulin Sensitivity and β-cell Glucose Sensitivity from the Response to Intravenous Glucose, *Journal of Clinical Investigation*, 68, 6, 1456–1467, 1981. doi:10.1172/JCI110398. 41

Bohez, S.; Abdolmaleki, A.; Neunert, M.; Buchli, J.; Heess, N.; and Hadsell, R., Value Constrained Model-Free Continuous Control, *arXiv:1902.04623*, 2019. doi:https://doi.org/10.48550/arXiv.1902.04623. 31

Bothe, M. K.; Dickens, L.; Reichel, K.; Tellmann, A.; Ellger, B.; Westphal, M.; and Faisal, A. A., The Use of Reinforcement Learning Algorithms to Meet the Challenges of an Artificial Pancreas, *Expert Review of Medical Devices*, 10, 5, 661–673, Taylor & Francis, 2013. doi:10.1586/17434440.2013.827515. 4, 5, 18, 25, 34, 124

Brazeau, A.; Mircescu, H.; Desjardins, K.; Leroux, C.; Strychar, I.; Ekoé, J.; and Rabasa-Lhoret, R., Carbohydrate Counting Accuracy and Blood Glucose Variability in Adults with Type 1 Diabetes, *Diabetes Research and Clinical Practice*, 99, 1, 19–23, Elsevier, 2013. doi:10.1016/j.diabres.2012.10.024. 2, 4, 22

Breton, M. D.; Brown, S. A.; Karvetski, C. H.; Kollar, L.; Topchyan, K. A.; Anderson, S. M.; and Kovatchev, B. P., Adding Heart Rate Signal to a Control-To-Range Artificial Pancreas System Improves the Protection Against Hypoglycemia During Exercise in Type 1 Diabetes, *Diabetes Technology and Therapeutics*, 16, 8, 506–511, 2014. doi:10.1089/dia.2013.0333. 24, 103, 104, 107, 142

Breton, M. D. and Kovatchev, B. P., One Year Real-World Use of the Control-IQ Advanced Hybrid Closed-Loop Technology, *Diabetes Technology & Therapeutics*, 9, 601–608, Mary Ann Liebert Inc, 2021. doi:10.1089/dia.2021.0097. 2

Brew-Sam, N.; Chhabra, M.; Parkinson, A.; Hannan, K.; Brown, E.; Pedley, L.; Brown, K.; Wright, K.; Pedley, E.; Nolan, C. J.; et al., Experiences of young people and their caregivers of using technology to manage type 1 diabetes mellitus: Systematic literature review and narrative synthesis, *JMIR Diabetes*, 6, 1, e20973, JMIR Publications Inc, 2021a. doi:10.2196/20973. 2, 4, 16, 21, 22, 109, 124, 125

Brew-Sam, N.; Chhabra, M.; Parkinson, A.; Henschke, A.; Brown, E.; Pedley, L.; Pedley, E.; Hannon, K.; Brown, K.; Wright, K.; et al., Diabetes Technology Experiences of Young People Living with Type 1 Diabetes Mellitus and their Parents: Hybrid Theoretical Foundation Guided Analysis, *MedRxiv*, 2021–11, Cold Spring Harbor Laboratory Press, 2021b. doi:https://doi.org/10.1101/2021.11.08.21265793. 4, 21, 39

Brew-Sam, N.; Parkinson, A.; Chhabra, M.; Henschke, A.; Brown, E.; Pedley, L.; Pedley, E.; Hannan, K.; Brown, K.; Wright, K.; et al., Toward Diabetes Device Development That Is Mindful to the Needs of Young People Living With Type 1 Diabetes: A Data-and Theory-Driven Qualitative Study, *JMIR Diabetes*, 8, 1, e43377, JMIR Publications Inc, 2023. doi:10.2196/43377. 4, 116

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W., Openai gym, *arXiv:1606.01540*, 2016. doi:https://doi.org/10.48550/arXiv.1606.01540. 87, 88, 94

Brunton, S. L. and Kutz, J. N., 2022. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press. 25

Cabitza, F.; Campagner, A.; and Sconfienza, L. M., Studying Human-AI Collaboration Protocols: The Case of the Kasparov's Law in Radiological Double Reading, *Health Information Science and Systems*, 9, 1–20, Springer, 2021. doi:10.1007/s13755-021-00138-8. 115

Campbell, M.; Hoane Jr, A. J.; and Hsu, F.-h., Deep Blue, *Artificial Intelligence*, 134, 1-2, 57–83, Elsevier, 2002. doi:https://doi.org/10.1016/S0004-3702(01)00129-1. 23, 29

Carolyn, J.; Lynn, G.; Ann, B.; Jo-Anne, M.-N.; Bjorn, N.; and Suneel, J., Key Research Summary - Parents Using Unregulated Technology to Manage Type 1 Diabetes in Children [Online], 2020. Available: https://law.unimelb.edu.au/__data/assets/pdf_file/0006/3506658/Personalised-Closed-Loop-Systems-for-Childhood-Diabetes-Research-Summary.pdf. [Accessed on 04/21/2023]. 2, 16, 116

CASP, CASP Checklists - CASP - Critical Appraisal Skills Programme [Online], 2020. Available: https://casp-uk.net/casp-tools-checklists/. [Accessed on 09/18/2020]. 98

Castle, J. R.; Youssef, J. E.; Wilson, L. M.; Reddy, R.; Resalat, N.; Branigan, D.; Ramsey, K.; Leitschuh, J.; Rajhbeharrysingh, U.; Senf, B.; Sugerman, S. M.;

Gabo, V.; and Jacobs, P. G., Randomized Outpatient Trial of Single- and Dual-Hormone Closed-Loop Systems that Adapt to Exercise Using Wearable Sensors, *Diabetes Care*, 41, 7, 1471–1477, 2018. doi:10.2337/dc18-0228. 24, 104, 107, 142

Chow, Y.; Nachum, O.; Duenez-Guzman, E.; and Ghavamzadeh, M., A Lyapunov-Based Approach to Safe Reinforcement Learning, *Advances in Neural Information Processing Systems*, 31, 2018. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/4fe5149039b52765bde64beb9f674940-Paper.pdf. 31

Chua, K.; Calandra, R.; McAllister, R.; and Levine, S., Deep Reinforcement Learning in a Handful of Trials Using Probabilistic Dynamics Models, *Advances in Neural Information Processing Systems*, 31, 2018. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/3de568f8597b94bda53149c7d7f5958c-Paper.pdf. 31

Cinar, A., Multivariable Adaptive Artificial Pancreas System in Type 1 Diabetes, *Current Diabetes Reports*, 17, 10, 1–11, Springer, 2017. doi:10.1007/s11892-017-0920-1. 3, 5, 20, 21, 96, 127

Cinar, A. and Oruklu, M., Automatic Insulin Pumps Using Recursive Multivariable Models and Adaptive Control Algorithms, 2014. US Patent 8,690,820. 108, 143

Cinar, A.; Turksoy, K.; and Hajizadeh, I., Multivariable Artificial Pancreas Method and System, 2020. US Patent 10,646,650. 108, 143

Cobbe, K. W.; Hilton, J.; Klimov, O.; and Schulman, J., Phasic Policy Gradient, *38th International Conference on Machine Learning*, Proceedings (Vol. 139 Pp. 2020–2027), PMLR, 2021. Available: https://proceedings.mlr.press/v139/cobbe21a.html. 33, 67

Cobelli, C.; Renard, E.; and Kovatchev, B., Artificial Pancreas: Past, Present, Future, *Diabetes*, 60, 11, 2672–2682, American Diabetes Association, 2011. doi:10.2337/db11-0654. 1, 15, 124

Colmegna, P.; Garelli, F.; De Battista, H.; and Sánchez-Peña, R., Automatic Regulatory Control in Type 1 Diabetes Without Carbohydrate Counting, *Control Engineering Practice*, 74, 22–32, Elsevier, 2018. doi:https://doi.org/10.1016/j.conengprac.2018.02.003. 18, 19, 20

Contreras, I. and Vehi, J., Artificial Intelligence for Diabetes Management and Decision Support: Literature Review, *Journal of Medical Internet Research*, 20, 5, e10775, JMIR Publications Inc, 2018. doi:10.2196/10775. 3, 23

Cui, R.; Hettiarachchi, C.; Nolan, C. J.; Daskalaki, E.; and Suominen, H., Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network, *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, Proceedings (Pp. 154–159), IEEE, 2021. doi:10.1109/CBMS52027.2021.00064. 120

Dalal, G.; Dvijotham, K.; Vecerik, M.; Hester, T.; Paduraru, C.; and Tassa, Y., Safe Exploration in Continuous Action Spaces, *arXiv:1801.08757*, 2018. doi:https://doi.org/10.48550/arXiv.1801.08757. 31

Dalla Man, C.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; and Cobelli, C., The UVA/PADOVA Type 1 Diabetes Simulator: New Features, *Journal of Diabetes Science and Technology*, 8, 1, 26–34, SAGE Publications, 2014. doi:10.1177/1932296813514502. 34, 42, 77

Daskalaki, E.; Diem, P.; and Mougiakakou, S. G., An Actor-Critic Based Controller for Glucose Regulation in Type 1 Diabetes, *Computer Methods and Programs in Biomedicine*, 109, 2, 116–125, Elsevier, 2013a. doi:10.1016/j.cmpb.2012.03.002. 25, 30, 34

Daskalaki, E.; Diem, P.; and Mougiakakou, S. G., Personalized Tuning of a Reinforcement Learning Control Algorithm for Glucose Regulation, *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3487–3490, IEEE, 2013b. doi:10.1109/EMBC.2013.6610293. 34

de la Vega, M.-L. L.; Fernandez-Mejia, C.; et al., Beta-cell Function and Failure in Type 1 Diabetes. *Type 1 Diabetes-Pathogenesis, Genetics and Immunotherapy*. IntechOpen, 2011. doi:10.5772/22089. 95

De Mol, C.; De Vito, E.; and Rosasco, L., Elastic-Net Regularization in Learning Theory, *Journal of Complexity*, 25, 2, 201–230, Elsevier, 2009. doi:https://doi.org/10.1016/j.jco.2009.01.002. 25

DeBoer, M. D.; Cherñavvsky, D. R.; Topchyan, K.; Kovatchev, B. P.; Francis, G. L.; and Breton, M. D., Heart Rate Informed Artificial Pancreas System Enhances Glycemic Control During Exercise in Adolescents with T1D, *Pediatric diabetes*, 18, 7, 540–546, Wiley Online Library, 2017. doi:10.1111/pedi.12454. 100, 103, 104, 106, 107, 110, 142

Desborough, J.; Parkinson, A.; Lewis, F.; Ebbeck, H.; Banfield, M.; and Phillips, C., A Framework for Involving Coproduction Partners in Research About Young People with Type 1 Diabetes, *Health Expectations*, 25, 1, 430–442, Wiley Online Library, 2022. doi:10.1111/hex.13403. 4, 116, 117

Diabetes Control and Complications Trial Research Group, The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus, *New England Journal of Medicine*, 329, 14, 977–986, Massachusetts Medical Society, 1993. doi:10.1056/NEJM199309303291401. 2, 3, 12, 13, 86

DiMeglio, L. A.; Evans-Molina, C.; and Oram, R. A., Type 1 diabetes, *The Lancet*, 391, 10138, 2449–2462, Elsevier, 2018. doi:10.1016/S0140-6736(18)31320-5. 1, 4, 11, 12, 64, 124

Dinneen, S. F. and McMorrow, L., Re-Framing Type 1 Diabetes Care Through Open-Source Automated Insulin Delivery: The (Expert) Patient Will See You Now, Doctor, *Diabetic Medicine*, 39, 5, Wiley-Blackwell, 2022. doi:10.1111/dme.14839. 116

D'Oro, P.; Metelli, A. M.; Tirinzoni, A.; Papini, M.; and Restelli, M., Gradient-Aware Model-Based Policy Search, *AAAI Conference on Artificial Intelligence*, Proceedings (Vol. 34, Pp. 3801–3808), 2020. doi:https://doi.org/10.1609/aaai.v34i04.5791. 128

Dudík, M.; Langford, J.; and Li, L., Doubly Robust Policy Evaluation and Learning, *arXiv:1103.4601*, 2011. doi:https://doi.org/10.48550/arXiv.1103.4601. 32

Dulac-Arnold, G.; Levine, N.; Mankowitz, D. J.; Li, J.; Paduraru, C.; Gowal, S.; and Hester, T., An Empirical Investigation of the Challenges of Real-World Reinforcement Learning, *arXiv:2003.11881*, 2020. doi:https://doi.org/10.48550/arXiv.2003.11881. 30, 71

Dulac-Arnold, G.; Mankowitz, D.; and Hester, T., Challenges of real-world reinforcement learning, *arXiv:1904.12901*, 2019. doi:https://doi.org/10.48550/arXiv.1904.12901. 4, 5, 30

Duriez, T.; Brunton, S. L.; and Noack, B. R., 2017. *Machine Learning Control - Taming Nonlinear Dynamics and Turbulence*, vol. 116. Springer. doi:https://doi.org/10.1007/978-3-319-40624-4. 25

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H., Hyperbolic Discounting and Learning Over Multiple Horizons, *arXiv:1902.06865*, 2019. doi:https://doi.org/10.48550/arXiv.1902.06865. 33, 63

Finn, C.; Abbeel, P.; and Levine, S., Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, *International conference on machine learning*, 70:1126–1135, PMLR, 2017. Available: https://proceedings.mlr.press/v70/finn17a.html. 32

Forlenza, G. P.; Ekhlaspour, L.; Breton, M.; Maahs, D. M.; Wadwa, R. P.; DeBoer, M.; Messer, L. H.; Town, M.; Pinnata, J.; Kruse, G.; et al., Successful At-Home Use of the Tandem Control-IQ Artificial Pancreas System in Young Children During a Randomized Controlled Trial, *Diabetes Technology & Therapeutics*, 21, 4, 159–169, Mary Ann Liebert, Inc, 2019. doi:10.1089/dia.2019.0011. 17

Fox, I.; Lee, J.; Pop-Busui, R.; and Wiens, J., Deep Reinforcement Learning for Closed-Loop Blood Glucose Control, *Machine Learning for Healthcare Conference*, 126:508–536, PMLR, 2020. Available: https://proceedings.mlr.press/v126/fox20a.html. 35, 36, 40, 45, 50, 51, 60, 61, 81, 82, 83, 87, 126

Fox, I. and Wiens, J., Reinforcement Learning for Blood Glucose Control: Challenges and Opportunities, 2019. Available: https://openreview.net/forum?id=ByexVzSAs4. 35, 36, 40, 60, 81, 82, 87, 126

Freeman, C. D.; Frey, E.; Raichuk, A.; Girgin, S.; Mordatch, I.; and Bachem, O., Brax–A Differentiable Physics Engine for Large Scale Rigid Body Simulation, *arXiv:2106.13281*, 2021. doi:https://doi.org/10.48550/arXiv.2106.13281. 119

Fritz, C. O.; Morris, P. E.; and Richler, J. J., Effect Size Estimates: Current Use, Calculations, and Interpretation, *Journal of Experimental Psychology: General*, 141, 1, 2, American Psychological Association, 2012. doi:10.1037/a0024338. 71

Gondhalekar, R.; Dassau, E.; Zisser, H. C.; and Doyle III, F. J., Periodic-Zone Model Predictive Control for Diurnal Closed-Loop Operation of an Artificial Pancreas, *Journal of Diabetes Science and Technology*, SAGE Publications, 2013. doi:10.1177/193229681300700605. 20, 21

Gottesman, O.; Johansson, F.; Komorowski, M.; Faisal, A.; Sontag, D.; Doshi-Velez, F.; and Celi, L. A., Guidelines for Reinforcement Learning in Healthcare, *Nature Medicine*, 25, 1, 16–18, Nature Publishing Group, 2019. doi:https://doi.org/10.1038/s41591-018-0310-5. 120

Granger, C. W., Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica: Journal of the Econometric Society*, 424–438, JSTOR, 1969. doi:https://doi.org/10.2307/1912791. 111

Gregory, G. A.; Robinson, T. I.; Linklater, S. E.; Wang, F.; Colagiuri, S.; de Beaufort, C.; Donaghue, K. C.; Magliano, D. J.; Maniam, J.; Orchard, T. J.; et al., Global Incidence, Prevalence, and Mortality of Type 1 Diabetes in 2021 with Projection to 2040: A Modelling Study, *The Lancet Diabetes & Endocrinology*, 10, 10, 741–760, Elsevier, 2022. doi:10.1016/S2213-8587(22)00218-2. 11, 12

Gu, S.; Lillicrap, T.; Sutskever, I.; and Levine, S., Continuous Deep Q-Learning with Model-Based Acceleration, *33rd International Conference on Machine Learning*, Proceedings (Vol. 48, Pp. 2829–2838), PMLR, 2016. Available: https://proceedings.mlr.press/v48/gu16.html. 29

Gu, S. S.; Lillicrap, T.; Turner, R. E.; Ghahramani, Z.; Schölkopf, B.; and Levine, S., Interpolated Policy Gradient: Merging On-Policy and Off-Policy Gradient Estimation for Deep Reinforcement Learning, *Advances in Neural Information Processing Systems*, 30, 2017. Available: https://dl.acm.org/doi/10.5555/3294996.3295141. 29

Güemes, A.; Herrero, P.; and Georgiou, P., A Novel Glucose Controller Using Insulin Sensitivity Modulation for Management of Type 1 Diabetes, *IEEE International Symposium on Circuits and Systems*, Proceedings (Pp. 1–5), IEEE Xplore, 2019. doi:10.1109/ISCAS.2019.8702535. 21

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al., Soft Actor-Critic Algorithms and Applications, *arXiv:1812.05905*, 2018. doi:https://doi.org/10.48550/arXiv.1812.05905. 4, 30, 32, 35, 40, 48, 49, 128

Hafner, D.; Lillicrap, T.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J., Learning Latent Dynamics for Planning From Pixels, *International Conference on Machine Learning*, Proceedings (Vol. 97, Pp. 2555–2565), PMLR, 2019. Available: https://proceedings.mlr.press/v97/hafner19a.html. 31

Hajizadeh, I.; Rashid, M.; Samadi, S.; Sevil, M.; Hobbs, N.; Brandt, R.; and Cinar, A., Adaptive Personalized Multivariable Artificial Pancreas Using Plasma Insulin Estimates, *Journal of Process Control*, 80, 26–40, Elsevier, 2019. doi:https://doi.org/10.1016/j.jprocont.2019.05.003. 24, 100, 102, 103, 105, 106, 142

Hajizadeh, I.; Rashid, M.; Turksoy, K.; Samadi, S.; Feng, J.; Sevil, M.; Frantz, N.; Lazaro, C.; Maloney, Z.; Littlejohn, E.; et al., Multivariable Recursive Subspace Identification with Application to Artificial Pancreas Systems, *IFAC-PapersOnLine*, 50, 1, 886–891, Elsevier, 2017. doi:https://doi.org/10.1016/j.ifacol.2017.08.268. 25

Hameed, H. and Kleinberg, S., Comparing Machine Learning Techniques for Blood Glucose Forecasting Using Free-Living and Patient Generated Data, *5th Machine Learning for Healthcare Conference*, Proceedings (Vol. 126, Pp. 871–894), PMLR, 2020. Available: https://proceedings.mlr.press/v126/hameed20a.html. 3, 23

Hansen, N.; Jangir, R.; Sun, Y.; Alenyà, G.; Abbeel, P.; Efros, A. A.; Pinto, L.; and Wang, X., Self-Supervised Policy Adaptation During Deployment, *arXiv:2007.04309*, 2020. doi:https://doi.org/10.48550/arXiv.2007.04309. 33

Hausknecht, M. and Stone, P., Deep Recurrent Q-Learning for Partially Observable MDPs, *AAAI Fall symposium Series*, AAAI Press, 2015. Available: https://aaai.org/papers/11673-deep-recurrent-q-learning-for-partially-observable-mdps/. 30

Haydari, A. and Yilmaz, Y., Deep Reinforcement Learning for Intelligent Transportation Systems: A Survey, *IEEE Transactions on Intelligent Transportation Systems*, 23, 1, 11–32, IEEE, 2020. doi:10.1109/TITS.2020.3008612. 4, 30

He, J.; Baxter, S. L.; Xu, J.; Xu, J.; Zhou, X.; and Zhang, K., The Practical Implementation of Artificial Intelligence Technologies in Medicine, *Nature Medicine*, 25, 1, 30–36, Nature Publishing Group, 2019. doi:https://doi.org/10.1038/s41591-018-0307-0. 116

Hernández-Ordoñez, M. and Campos-Delgado, D. U., An Extension to the Compartmental Model of Type 1 Diabetic Patients to Reproduce Exercise Periods with Glycogen Depletion and Replenishment, *Journal of Biomechanics*, 41, 4, 744–752, 2008. doi:10.1016/j.jbiomech.2007.11.028. 103, 105

Hessel, M.; Danihelka, I.; Viola, F.; Guez, A.; Schmitt, S.; Sifre, L.; Weber, T.; Silver, D.; and Van Hasselt, H., Muesli: Combining Improvements in Policy Optimization, *International Conference on Machine Learning*, Proceedings (Vol. 139, Pp. 4214–4226), PMLR, 2021. Available: https://proceedings.mlr.press/v139/hessel21a.html. 33

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al., Deep Q-Learning from Demonstrations, *AAAI Conference on Artificial Intelligence*, Proceedings (Vol. 32, No. 1), 2018. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11757/11616. 30

Hettiarachchi, C.; Daskalaki, E.; Desborough, J.; Nolan, C. J.; O'Neal, D.; Suominen, H.; et al., Integrating Multiple Inputs Into an Artificial Pancreas System: Narrative Literature Review, *JMIR Diabetes*, 7, 1, e28861, JMIR Publications, 2022a. doi:10.2196/28861. 24

Hettiarachchi, C.; Malagutti, N.; Nolan, C.; Daskalaki, E.; and Suominen, H., A Reinforcement Learning Based System for Blood Glucose Control without Carbohydrate Estimation in Type 1 Diabetes: In Silico Validation, *35th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 950–956, IEEE, 2022b. doi:10.1109/EMBC48229.2022.9871054. 45, 52, 81, 82, 83

Hihi, S. and Bengio, Y., Hierarchical Recurrent Neural Networks for Long-Term Dependencies, *Advances in Neural Information Processing Systems*, 8, 1995. Available: https://proceedings.neurips.cc/paper/1995/file/c667d53acd899a97a85de0c201ba99be-Paper.pdf. 23

Hochreiter, S. and Schmidhuber, J., Long Short-Term Memory, *Neural Computation*, 9, 8, 1735–1780, MIT Press, 1997. doi:https://doi.org/10.1162/neco.1997.9.8.1735. 23, 25, 65

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.; and Wilinska, M. E., Nonlinear Model Predictive Control of Glucose Concentration in Subjects with Type 1 Diabetes, *Physiological Measurement*, 25, 4, 905–920, 2004. doi:10.1088/0967-3334/25/4/010. 103, 105

Hughes, A. S.; Bispham, J.; Fan, L.; Nieves-Perez, M.; and McAuliffe-Fogarty, A. H., "I Live in Constant Fear of Highs," the Daily Impact of Type 1 Diabetes, *Journal of Patient Experience*, 7, 6, 911–914, SAGE Publications, 2020. doi:10.1177/2374373520967501. 39

Hung, C.-C.; Lillicrap, T.; Abramson, J.; Wu, Y.; Mirza, M.; Carnevale, F.; Ahuja, A.; and Wayne, G., Optimizing Agent Behavior Over Long Time Scales by Transporting Value, *Nature Communications*, 10, 1, 5223, Nature Publishing Group, 2019. doi:https://doi.org/10.1038/s41467-019-13073-w. 30

Insulin Pump Comparison, 2022 [Online], Insulin Pump Comparison, *Accessed on: March 24th, 2022*. Available: https://www.betterlivingnow.com/forms/Insulin-Pump-Comparison.pdf. 86

Iqbal, M. H.; Aydin, A.; Brunckhorst, O.; Dasgupta, P.; and Ahmed, K., A Review of Wearable Technology in Medicine, *Journal of the Royal Society of Medicine*, 109, 10, 372–380, SAGE Publications, 2016. doi:10.1177/0141076816663560. 5, 110

Iturralde, E.; Tanenbaum, M. L.; Hanes, S. J.; Suttiratana, S. C.; Ambrosino, J. M.; Ly, T. T.; Maahs, D. M.; Naranjo, D.; Walders-Abramson, N.; Weinzimer, S. A.; et al., Expectations and Attitudes of Individuals with Type 1 Diabetes After Using a Hybrid Closed Loop System, *The Diabetes Educator*, 43, 2, 223–232, Sage Publications, 2017. doi:10.1177/0145721717697244. 109

Jacobs, P. G.; El Youssef, J.; Reddy, R.; Resalat, N.; Branigan, D.; Condon, J.; Preiser, N.; Ramsey, K.; Jones, M.; Edwards, C.; Kuehl, K.; Leitschuh, J.; Rajhbeharrysingh, U.; and Castle, J. R., Randomized Trial of a Dual-Hormone Artificial Pancreas with Dosing Adjustment During Exercise Compared with no Adjustment and Sensor-Augmented Pump Therapy, *Diabetes, Obesity and Metabolism*, 18, 11, 1110–1119, Blackwell Publishing Ltd, 2016. doi:10.1111/dom.12707. 103, 104, 107, 142

Jacobs, P. G.; Resalat, N.; El Youssef, J.; Reddy, R.; Branigan, D.; Preiser, N.; Condon, J.; and Castle, J., Incorporating an Exercise Detection, Grading, and Hormone Dosing Algorithm into the Artificial Pancreas Using Accelerometry and Heart Rate, *Journal of Diabetes Science and Technology*, 9, 6, 1175–1184, SAGE Publications, 2015. doi:10.1177/1932296815609371. 100, 102, 103, 105, 106, 142

Janner, M.; Fu, J.; Zhang, M.; and Levine, S., When to Trust Your Model: Model-Based Policy Optimization, *Advances in Neural Information Processing Systems*, 32, 2019. Available: https://proceedings.neurips.cc/paper/2019/file/5faf461eff3099671ad63c6f3f094f7f-Paper.pdf. 128

Jiang, N. and Li, L., Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning, *International Conference on Machine Learning*, Proceedings (Vol. 48, Pp. 652–661), PMLR, 2016. Available: https://proceedings.mlr.press/v48/jiang16.html. 32

Jordan, M. I. and Mitchell, T. M., Machine Learning: Trends, Perspectives, and Prospects, *Science*, 349, 6245, 255–260, American Association for the Advancement of Science, 2015. Available: https://www.science.org/doi/pdf/10.1126/science.aaa8415. 3

Kalra, S.; Mukherjee, J. J.; Venkataraman, S.; Bantwal, G.; Shaikh, S.; Saboo, B.; Das, A. K.; and Ramachandran, A., Hypoglycemia: The Neglected Complication, *Indian Journal of Endocrinology and Metabolism*, 17, 5, 819, Wolters Kluwer–Medknow Publications, 2013. doi:10.4103/2230-8210.117219. 51, 83

Kasparov, G., 2017. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. Hachette, UK. 115

Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; and Chouvarda, I., Machine Learning and Data Mining Methods in Diabetes Research, *Computational and Structural Biotechnology Journal*, 15, 104–116, Elsevier, 2017. doi:10.1016/j.csbj.2016.12.005. 3, 23

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, Proceedings (Vol. 30, Pp. 3146–3154), Curran Associates Inc., 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. 25

Kendall, A.; Hawke, J.; Janz, D.; Mazur, P.; Reda, D.; Allen, J.-M.; Lam, V.-D.; Bewley, A.; and Shah, A., Learning to Drive in a Day, *International Conference on Robotics and Automation (ICRA)*, Proceedings (Pp. 8248–8254), IEEE, 2019. doi:10.1109/ICRA.2019.8793742. 30, 32

Kesavadev, J.; Srinivasan, S.; Saboo, B.; Krishna, B.; and Krishnan, G., The Do-It-Yourself Artificial Pancreas: A Comprehensive Review, *Diabetes Therapy: Research, Treatment and Education of Diabetes and Related Disorders*, 11, 6, 1217–1235, Springer, 2020. doi:10.1007/s13300-020-00823-z. 4, 116

Khan, S. H.; Khan, A. H.; and Khan, Z. H., Artificial Pancreas Coupled Vital Signs Monitoring for Improved Patient Safety, *Arabian Journal for Science and Engineering*, 38, 11, 3093–3102, Springer, 2013. doi:https://doi.org/10.1007/s13369-012-0456-2. 100, 103, 105, 106, 142

Khorasgani, H.; Zhang, C.; Gupta, C.; and Serita, S., Long-Term Planning, Short-Term Adjustments, 2019. Available: https://openreview.net/pdf?id=SJlbyCNtPr. 33, 63

Kingma, D. P. and Ba, J., Adam: A Method for Stochastic Optimization, *arXiv:1412.6980*, 2014. doi:https://doi.org/10.48550/arXiv.1412.6980. 49, 50

Kogiso, K. and Fujita, T., Cyber-Security Enhancement of Networked Control Systems Using Homomorphic Encryption, *54th IEEE Conference on Decision and Control (CDC)*, Proceedings (Pp. 6836–6843), IEEE, 2015. doi:10.1109/CDC.2015.7403296. 121

Kovatchev, B. P.; Breton, M.; Dalla Man, C.; and Cobelli, C., 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, SAGE Publications, 2009. doi:10.1177/193229680900300106. 16, 20, 21, 36, 37, 41, 42, 60, 64, 83, 119, 127

Kovatchev, B. P.; Clarke, W. L.; Breton, M.; Brayman, K.; and McCall, A., Quantifying Temporal Glucose Variability in Diabetes via Continuous Glucose Monitoring: Mathematical Methods and Clinical Application, *Diabetes Technology and Therapeutics*, 7, 6, 849–862, Mary Ann Liebert, Inc, 2005. doi:10.1089/dia.2005.7.849. 45, 51

Koza, J. R.; Bennett, F. H.; and Stiffelman, O., Genetic Programming as a Darwinian Invention Machine, *Genetic Programming: Second European Workshop*, Proceedings (Vol. 2, Pp. 93–108), Springer, 1999. doi:https://doi.org/10.1007/3-540-48885-5_8. 25

Kreisman, S. H.; Ah Mew, N.; Halter, J. B.; Vranic, M.; and Marliss, E. B., Norepinephrine Infusion During Moderate-Intensity Exercise Increases Glucose Production and Uptake, *Journal of Clinical Endocrinology and Metabolism*, 86, 5, 2118–2124, 2001. doi:10.1210/jc.86.5.2118. 103, 105

Kudva, Y. C.; Carter, R. E.; Cobelli, C.; Basu, R.; and Basu, A., Closed-Loop Artificial Pancreas Systems: Physiological Input to Enhance Next-Generation Devices, *Diabetes Care*, 37, 5, 1184–1190, American Diabetes Association Inc., 2014. doi:10.2337/dc13-2066. 96, 127

Kurutach, T.; Clavera, I.; Duan, Y.; Tamar, A.; and Abbeel, P., Model-Ensemble Trust-Region Policy Optimization, *arXiv:1802.10592*, 2018. doi:https://doi.org/10.48550/arXiv.1802.10592. 128

Lan, T.; Srinivasa, S.; Wang, H.; and Zheng, S., Warpdrive: Extremely Fast End-To-End Deep Multi-Agent Reinforcement Learning on a GPU, *arXiv:2108.13976*, 2021. doi:https://doi.org/10.48550/arXiv.2108.13976. 119

Lawton, J.; Blackburn, M.; Allen, J.; Campbell, F.; Elleri, D.; Leelarathna, L.; Rankin, D.; Tauschmann, M.; Thabit, H.; and Hovorka, R., Patients' and Caregivers' Experiences of Using Continuous Glucose Monitoring to Support Diabetes Self-Management: Qualitative Study, *BMC Endocrine Disorders*, 18, 1, 1–10, BioMed Central, 2018. doi:10.1186/s12902-018-0239-1. 109

Lazaric, A.; Restelli, M.; and Bonarini, A., Reinforcement Learning in Continuous Action Spaces Through Sequential Monte Carlo Methods, *Advances in Neural Information Processing Systems*, 20, 2007. Available: https://proceedings.neurips.cc/paper_files/paper/2007/file/0f840be9b8db4d3fbd5ba2ce59211f55-Paper.pdf. 31, 86, 94

Lea, C.; Vidal, R.; Reiter, A.; and Hager, G. D., Temporal Convolutional Networks: A Unified Approach to Action Segmentation, *European Conference on Computer Vision*, 47–54, Springer, 2016. doi:https://doi.org/10.1007/978-3-319-49409-8_7. 23, 25

LeCun, Y.; Bengio, Y.; and Hinton, G., Deep Learning, *Nature*, 521, 7553, 436–444, Nature Publishing Group, 2015. doi:https://doi.org/10.1038/nature14539. 3, 23

LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; and Jackel, L., Handwritten Digit Recognition With a Back-Propagation Network, *Advances in Neural Information Processing Systems*, 2, 1989. Available: https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf. 23

Lee, A. X.; Nagabandi, A.; Abbeel, P.; and Levine, S., Stochastic Latent Actor-Critic: Deep Reinforcement Learning With a Latent Variable Model, *Advances in Neural Information Processing Systems*, 33, 741–752, 2020a. Available: https://proceedings.neurips.cc/paper/2020/file/08058bf500242562c0d031ff830ad094-Paper.pdf. 33

Lee, M. H.; Paldus, B.; Krishnamurthy, B.; McAuley, S. A.; Shah, R.; Jenkins, A. J.; and O'Neal, D. N., The Clinical Case for the Integration of a Ketone Sensor as Part of a Closed Loop Insulin Pump System, *Journal of Diabetes Science and Technology*, 13, 5, 967–973, 2019. doi:10.1177/1932296818822986. 24, 109, 110

Lee, S.; Kim, J.; Park, S. W.; Jin, S.-M.; and Park, S.-M., Toward a Fully Automated Artificial Pancreas System Using a Bioinspired Reinforcement Learning Design: In Silico Validation, *IEEE Journal of Biomedical and Health Informatics*, 25, 2, 536–546, IEEE, 2020b. doi:10.1109/JBHI.2020.3002022. 35, 36, 40, 46, 52, 60, 61, 81, 82, 83, 87, 126, 127

Leelarathna, L.; Choudhary, P.; Wilmot, E. G.; Lumb, A.; Street, T.; Kar, P.; and Ng, S. M., Hybrid Closed-Loop Therapy: Where Are We In 2021?, *Diabetes, Obesity and Metabolism*, 23, 3, 655–660, Wiley Online Library, 2021. doi:10.1111/dom.14273. 2

Leelarathna, L.; Thabit, H.; Willinska, M. E.; Bally, L.; Mader, J. K.; Arnolds, S.; Benesch, C.; Pieber, T. R.; Shah, V. N.; Carlson, A. L.; et al., Duration of Hybrid Closed-Loop Insulin Therapy to Achieve Representative Glycemic Outcomes in Adults With Type 1 Diabetes, *Diabetes Care*, 43, 3, e38–e39, American Diabetes Association, 2020. doi:10.2337/dc19-2041. 17

Li, L., MATLAB User Manual, *The MathWorks, Inc., Natick, MA, USA*, 2001. 41, 81

Liaw, A.; Wiener, M.; et al., Classification and Regression by RandomForest, *R news*, 2, 3, 18–22, 2002. Available: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf. 25

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D., Continuous Control with Deep Reinforcement Learning, *arXiv:1509.02971*, 2015. doi:https://doi.org/10.48550/arXiv.1509.02971. 30, 32, 86

Lim, M. H.; Lee, W. H.; Jeon, B.; and Kim, S., A Blood Glucose Control Framework Based on Reinforcement Learning With Safety and Interpretability: In Silico Validation, *IEEE Access*, 9, 105756–105775, IEEE, 2021. doi:10.1109/ACCESS.2021.3100007. 35, 36, 61, 81, 82, 83, 87, 127

Liu, S.; See, K. C.; Ngiam, K. Y.; Celi, L. A.; Sun, X.; Feng, M.; et al., Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review, *Journal of Medical Internet Research*, 22, 7, e18477, JMIR Publications Inc., 2020. doi:10.2196/18477. 4, 30

Lizier, J. and Rubinov, M., Multivariate Construction of Effective Computational Networks from Observational Data, 2012. Available: https://www.mis.mpg.de/preprints/2012/preprint2012_25.pdf. 111

Ljung, L.; Andersson, C.; Tiels, K.; and Schön, T. B., Deep Learning and System Identification, *IFAC-PapersOnLine*, 53, 2, 1175–1181, Elsevier, 2020. doi:https://doi.org/10.1016/j.ifacol.2020.12.1329. 67

Luo, J.; Paduraru, C.; Voicu, O.; Chervonyi, Y.; Munns, S.; Li, J.; Qian, C.; Dutta, P.; Davis, J. Q.; Wu, N.; et al., Controlling Commercial Cooling Systems Using Reinforcement Learning, *arXiv:2211.07357*, 2022. doi:https://doi.org/10.48550/arXiv.2211.07357. 4, 30

Lv, D.; Breton, M. D.; and Farhy, L. S., Pharmacokinetics Modeling of Exogenous Glucagon in Type 1 Diabetes Mellitus Patients, *Diabetes Technology and Therapeutics*, 15, 11, 935–941, 2013. doi:10.1089/dia.2013.0150. 103, 105

Makhni, S., Co-Creation in Health Systems Design, *AMA Journal of Ethics*, 19, 11, 1070–1072, American Medical Association, 2017. doi:10.1001/journalofethics.2017.19.11.fred1-1711. 116

Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; et al., Isaac Gym: High Performance GPU-Based Physics Simulation for Robot Learning, *arXiv:2108.10470*, 2021. doi:https://doi.org/10.48550/arXiv.2108.10470. 119

Mankowitz, D. J.; Tamar, A.; and Mannor, S., Situational Awareness by Risk-Conscious Skills, *arXiv:1610.02847*, 2016. doi:https://doi.org/10.48550/arXiv.1610.02847. 31

Mann, H. B. and Whitney, D. R., On a Test of Whether One of Two Random Variables is Stochastically Larger Than the Other, *The Annals of Mathematical Statistics*, 50–60, JSTOR, 1947. Available: https://www.jstor.org/stable/2236101. 71, 90

Marling, C. and Bunescu, R., The OhioT1DM dataset for Blood Glucose Level Prediction: Update 2020, *CEUR Workshop*, Proceedings (Vol. 2675, Pp. 71, NIH Public Access, 2020. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7881904/. 120

McCarthy, J.; Minsky, M. L.; Rochester, N.; and Shannon, C. E., A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955, *AI Magazine*, 27, 4, 12–12, 2006. doi:https://doi.org/10.1609/aimag.v27i4.1904. 22

Medtronic, MiniMed™ 780G Insulin Pump System [Online], 2019. Available: https://hcp.medtronic-diabetes.com.au/sites/default/files/minimed_780g_hcp_brochure.pdf. [Accessed on 04/19/2023]. 22

Messer, L. H.; Forlenza, G. P.; Sherr, J. L.; Wadwa, R. P.; Buckingham, B. A.; Weinzimer, S. A.; Maahs, D. M.; and Slover, R. H., Optimizing Hybrid Closed-Loop Therapy in Adolescents and Emerging Adults Using the MiniMed 670G System, *Diabetes Care*, 41, 4, 789–796, American Diabetes Association, 2018. doi:10.2337/dc17-1682. 17

Minsky, M., Steps Toward Artificial Intelligence, *Proceedings of the IRE*, 49, 1, 8–30, IEEE, 1961. doi:10.1109/JRPROC.1961.287775. 3

Minsky, M., 1968. *Semantic Information Processing*. MIT Press, Cambridge, Massachusetts. 22

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K., Asynchronous Methods for Deep Reinforcement Learning, *33rd International Conference on Machine Learning*, Proceedings (Vol. 48, Pp. 1928–1937), PMLR, 2016. Available: http://proceedings.mlr.press/v48/mniha16.pdf. 32, 40, 128

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M., Playing Atari with Deep Reinforcement Learning, *arXiv:1312.5602*, 2013. doi:https://doi.org/10.48550/arXiv.1312.5602. 4, 30, 32

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al., Human-Level Control Through Deep Reinforcement Learning, *nature*, 518, 7540, 529–533, Nature Publishing Group, 2015. doi:https://doi.org/10.1038/nature14236. 30, 32

Myhre, J. N.; Launonen, I. K.; Wei, S.; and Godtliebsen, F., Controlling Blood Glucose Levels in Patients with Type 1 Diabetes Using Fitted Q-Iterations and Functional Features, *IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6, IEEE, 2018. doi:10.1109/MLSP.2018.8516946. 35, 87, 126

Nagabandi, A.; Finn, C.; and Levine, S., Deep Online Learning via Meta-Learning: Continual Adaptation for Model-Based RL, *arXiv:1812.07671*, 2018a. doi:https://doi.org/10.48550/arXiv.1812.07671. 30

Nagabandi, A.; Kahn, G.; Fearing, R. S.; and Levine, S., Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning, *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 7559–7566, IEEE, 2018b. doi:10.1109/ICRA.2018.8463189. 32

Naik, A.; Shariff, R.; et al., Discounted Reinforcement Learning Is Not an Optimization Problem, *arXiv:1910.02140*, 2019. doi:https://doi.org/10.48550/arXiv.1910.02140. 46, 47

Nathan, D.; Genuth, S.; et al., The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus, *The New England Journal of Medicine*, 329, 14, 977–986, 1993. doi:10.1056/NEJM199309303291401. 13, 14, 21, 60, 64

Ng, A. Y.; Russell, S.; et al., Algorithms for Inverse Reinforcement Learning, *International Conference on Machine Learning*, Proceedings (Vol. 1, Pp. 2), 2000. Available: https://ai.stanford.edu/~ang/papers/icml00-irl.pdf. 31

Ngo, P. D.; Wei, S.; Holubová, A.; Muzik, J.; and Godtliebsen, F., Control of Blood Glucose for Type-1 Diabetes by Using Reinforcement Learning With Feedforward Algorithm, *Computational and Mathematical Methods in Medicine*, 2018:4091497, Hindawi, 2018a. doi:10.1155/2018/4091497. 35, 60, 87, 126

Ngo, P. D.; Wei, S.; Holubová, A.; Muzik, J.; and Godtliebsen, F., Reinforcement-Learning Optimal Control for Type-1 Diabetes, *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Proceedings (Pp. 333–336), IEEE, 2018b. doi:10.1109/BHI.2018.8333436. 35, 60, 87, 126

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al., Training Language Models to Follow Instructions With Human Feedback, *arXiv:2203.02155*, 2022. doi:https://doi.org/10.48550/arXiv.2203.02155. 30

Paldus, B.; Morrison, D.; Lee, M.; Zaharieva, D. P.; Riddell, M. C.; and O'Neal, D. N., Strengths and Challenges of Closed-Loop Insulin Delivery During Exercise in People With Type 1 Diabetes: Potential Future Directions, *Journal of Diabetes Science and Technology*, 19322968221088327, SAGE Publications, 2022a. doi:10.1177/19322968221088327. 110

Paldus, B.; Morrison, D.; Zaharieva, D. P.; Lee, M. H.; Jones, H.; Obeyesekere, V.; Lu, J.; Vogrin, S.; La Gerche, A.; McAuley, S. A.; et al., A Randomized Crossover Trial Comparing Glucose Control During Moderate-Intensity, High-Intensity, and Resistance Exercise With Hybrid Closed-Loop Insulin Delivery While Profiling Potential Additional Signals in Adults With Type 1 Diabetes, *Diabetes care*, 45, 1, 194–203, American Diabetes Association, 2022b. doi:10.2337/dc21-1593. 1, 110, 111, 112, 113, 127

Palnitkar, R. M. and Cannady, J., A Review of Adaptive Neural Networks, *IEEE SoutheastCon*, Proceedings (Pp. 38–47), IEEE, 2004. doi:10.1109/SECON.2004.1287896. 25

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, *Advances in Neural Information Processing Systems*, 32, 8026–8037, 2019. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf. 35, 42, 49

Patek, S. D., Multiple-Signal Artificial Pancreas Systems. *The Artificial Pancreas*, 219–235. Elsevier. ISBN 9780128156551, 2019. doi:10.1016/b978-0-12-815655-1.00019-3. 96, 127

Perolat, J.; De Vylder, B.; Hennes, D.; Tarassov, E.; Strub, F.; de Boer, V.; Muller, P.; Connor, J. T.; Burch, N.; Anthony, T.; et al., Mastering the Game of Stratego with Model-Free Multiagent Reinforcement Learning, *Science*, 378, 6623, 990–996, American Association for the Advancement of Science, 2022. doi:10.1126/science.add4679. 30

Peters, T. M. and Haidar, A., Dual-Hormone Artificial Pancreas: Benefits and Limitations Compared with Single-Hormone Systems, *Diabetic Medicine*, 35, 4, 450–459, 2018. doi:10.1111/dme.13581. 21

Petersen, B. K.; Yang, J.; Grathwohl, W. S.; Cockrell, C.; Santiago, C.; An, G.; and Faissol, D. M., Precision Medicine as a Control Problem: Using Simulation and Deep Reinforcement Learning to Discover Adaptive, Personalized Multi-Cytokine Therapy for Sepsis, *arXiv:1802.10440*, 2018. doi:https://doi.org/10.48550/arXiv.1802.10440. 30

Petrovski, G.; Al Khalaf, F.; Campbell, J.; Umer, F.; Almajaly, D.; Hamdan, M.; and Hussain, K., One-Year Experience of Hybrid Closed-Loop System in Children and Adolescents With Type 1 Diabetes Previously Treated with Multiple Daily Injections: Drivers to Successful Outcomes, *Acta diabetologica*, 58, 207–213, Springer, 2021. doi:10.1007/s00592-020-01607-4. 17

Pham, T.-H.; De Magistris, G.; and Tachibana, R., Optlayer-Practical Constrained Optimization for Deep Reinforcement Learning in the Real World, *IEEE International Conference on Robotics and Automation (ICRA)*, 6236–6243, IEEE, 2018. doi:10.1109/ICRA.2018.8460547. 31

Precup, D.; Richard, S.; and Singh, S., Eligibility Traces for Off-Policy Policy Evaluation, *7th International Conference on Machine Learning*, Proceedings (Pp. 759–766), 2000. Available: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1079&context=cs_faculty_pubs. 32

PRISMA, PRISMA: Transparent Reporting of Systematic Reviews and Meta-Analysis [Online], 2020. Available: http://www.prisma-statement.org/. (Accessed on 09/18/2020). 97

Qaisar, S. B.; Khan, S. H.; and Imtiaz, S., Neural network and physiological parameters based control of artificial pancreas for improved patient safety, *International Conference on Computational Science and Its Applications*, Proceedings (Vol. 7335, Pp. 339–351), Springer, 2012. doi:https://doi.org/10.1007/978-3-642-31137-6_26. 100, 103, 105, 106, 142

Quiroz, G. and Femat, R., Theoretical Blood Glucose Control in Hyper-and Hypoglycemic and Exercise Scenarios by Means of an H_∞ Algorithm, *Journal of Theoretical Biology*, 263, 1, 154–160, Elsevier, 2010. doi:10.1016/j.jtbi.2009.11.015. 100, 102, 103, 105, 106, 113, 142

Quiroz, G.; Flores-Gutiérrez, C.; and Femat, R., Suboptimal H_∞ Hyperglycemia Control on T1DM Accounting Biosignals of Exercise and Nocturnal Hypoglycemia, *Optimal Control Applications and Methods*, 32, 2, 239–252, Wiley Online Library, 2011. doi:https://doi.org/10.1002/oca.989. 100, 102, 103, 105, 106, 113, 142

Ramkissoon, C. M.; Herrero, P.; Bondia, J.; and Vehi, J., Unannounced Meals in the Artificial Pancreas: Detection Using Continuous Glucose Monitoring, *Sensors*, 18, 3, 884, MDPI, 2018. doi:10.3390/s18030884. 24, 103

Rankin, D.; Harden, J.; Barnard, K.; Bath, L.; Noyes, K.; Stephen, J.; and Lawton, J., Barriers and Facilitators to Taking on Diabetes Self-Management Tasks in Pre-Adolescent Children with Type 1 Diabetes: A Qualitative Study, *BMC Endocrine Disorders*, 18, 1, 1–9, BioMed Central, 2018. doi:10.1186/s12902-018-0302-y. 109

Rashid, M.; Samadi, S.; Sevil, M.; Hajizadeh, I.; Kolodziej, P.; Hobbs, N.; Maloney, Z.; Brandt, R.; Feng, J.; Park, M.; Quinn, L.; and Cinar, A., Simulation Software for Assessment of Nonlinear and Adaptive Multivariable Control Algorithms: Glucose–Insulin Dynamics in Type 1 Diabetes, *Computers and Chemical Engineering*, 130, Elsevier, 2019. doi:10.1016/j.compchemeng.2019.106565. 25, 41, 103, 105

Resalat, N.; El Youssef, J.; Tyler, N.; Castle, J.; and Jacobs, P. G., A statistical virtual patient population for the glucoregulatory system in type 1 diabetes with integrated exercise model, *PLOS ONE*, 14, 7, e0217301, 2019a. doi:10.1371/journal.pone.0217301. Available: http://dx.plos.org/10.1371/journal.pone.0217301. 41, 103, 105

Resalat, N.; Hilts, W.; Youssef, J. E.; Tyler, N.; Castle, J. R.; and Jacobs, P. G., Adaptive Control of an Artificial Pancreas Using Model Identification, Adaptive Postprandial Insulin Delivery, and Heart Rate and Accelerometry as Control Inputs, *Journal of Diabetes Science and Technology*, 13, 6, 1044–1053, SAGE Publications, 2019b. doi:10.1177/1932296819881467. 100, 102, 103, 105, 106, 142

Roep, B. O., Improving Clinical Islet Transplantation Outcomes, *Diabetes Care*, 43, 4, 698–700, American Diabetes Association, 2020. doi:10.2337/dci19-0080. 13

Romoff, J.; Henderson, P.; Touati, A.; Brunskill, E.; Pineau, J.; and Ollivier, Y., Separating Value Functions Across Time-Scales, *International Conference on Machine Learning*, Proceedings (Vol. 97, Pp. 5468–5477), PMLR, 2019. Available: http://proceedings.mlr.press/v97/romoff19a/romoff19a.pdf. 33, 63

Rorsman, P.; Eliasson, L.; Renstrom, E.; Gromada, J.; Barg, S.; and Gopel, S., The Cell Physiology of Biphasic Insulin Secretion, *Physiology*, 15, 2, 72–77, American Physiological Society, 2000. doi:10.1152/physiologyonline.2000.15.2.72. 1

Roversi, C.; Vettoretti, M.; Del Favero, S.; Facchinetti, A.; Sparacino, G.; and Consortium, H.-R., Modeling Carbohydrate Counting Error in Type 1 Di-

abetes Management, *Diabetes Technology and Therapeutics*, 22, 10, 749–759, Mary Ann Liebert, Inc., 2020. doi:10.1089/dia.2019.0502. 22, 50

Sadler, M. and Regan, N., 2019. *Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI*. New in Chess, Alkmaar, The Netherlands. 27

Samuel, A. L., Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal of Research and Development*, 3, 3, 210–229, IBM, 1959. doi:10.1147/rd.33.0210. 3, 28

Sanchez, E. N., 2019. *The Artificial Pancreas: Current Situation and Future Directions*. Academic Press. 41

Saunders, A.; Messer, L. H.; and Forlenza, G. P., MiniMed 670G Hybrid Closed Loop Artificial Pancreas System for the Treatment of Type 1 Diabetes Mellitus: Overview of its Safety and Efficacy, *Expert Review of Medical Devices*, 16, 10, 845–853, Taylor and Francis, 2019. doi:10.1080/17434440.2019.1670639. 2

Saunders, W.; Sastry, G.; Stuhlmueller, A.; and Evans, O., Trial Without Error: Towards Safe Reinforcement Learning via Human Intervention, *arXiv:1707.05173*, 2017. doi:https://doi.org/10.48550/arXiv.1707.05173. 31

Schreiber, T., Measuring Information Transfer, *Physical Review Letters*, 85, 2, 461, American Physical Society, 2000. doi:https://doi.org/10.1103/PhysRevLett.85.461. 111

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al., Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model, *Nature*, 588, 7839, 604–609, Nature Publishing Group, 2020. doi:https://doi.org/10.1038/s41586-020-03051-4. 4, 29, 30, 33, 128

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O., Proximal Policy Optimization Algorithms, *arXiv:1707.06347*, 2017. doi:https://doi.org/10.48550/arXiv.1707.06347. 4, 30, 32, 35, 40, 48, 66, 80, 128

Schultes, B.; Jauch-Chara, K.; Gais, S.; Hallschmid, M.; Reiprich, E.; Kern, W.; Oltmanns, K. M.; Peters, A.; Fehm, H. L.; and Born, J., Defective Awakening Response to Nocturnal Hypoglycemia in Patients with Type 1 Diabetes Mellitus, *PLOS Medicine*, 4, 2, 0361–0369, 2007. doi:10.1371/journal.pmed.0040069. 103, 105

Schwarzer, M.; Anand, A.; Goel, R.; Hjelm, R. D.; Courville, A.; and Bachman, P., Data-Efficient Reinforcement Learning with Self-Predictive Representations, *arXiv:2007.05929*, 2020. doi:https://doi.org/10.48550/arXiv.2007.05929. 33

Shah, R. B.; Patel, M.; Maahs, D. M.; and Shah, V. N., Insulin Delivery Methods: Past, Present and Future, *International Journal of Pharmaceutical Investigation*, 6, 1, 1, Wolters Kluwer–Medknow Publications, 2016. doi:10.4103/2230-973X.176456. 13, 22

Shapiro, S. S. and Wilk, M. B., An Analysis of Variance Test for Normality (Complete Samples), *Biometrika*, 52, 3/4, 591–611, JSTOR, 1965. doi:https://doi.org/10.2307/2333709. 71, 89

Sherwood, J. S.; Russell, S. J.; and Putman, M. S., New and Emerging Technologies in Type 1 Diabetes, *Endocrinology and Metabolism Clinics*, 49, 4, 667–678, Elsevier, 2020. doi:10.1016/j.ecl.2020.07.006. 2, 16

Silva, J. D.; Lepore, G.; Battelino, T.; Arrieta, A.; Castañeda, J.; Grossman, B.; Shin, J.; and Cohen, O., Real-World Performance of the MiniMed^{TM} 780G System: First Report of Outcomes From 4120 Users, *Diabetes Technology & Therapeutics*, 24, 2, 113–119, Mary Ann Liebert, Inc., 2022. doi:10.1089/dia.2021.0203. 17

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*, 529, 7587, 484–489, Nature Publishing Group, 2016. doi:https://doi.org/10.1038/nature16961. 4, 30

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al., Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, *arXiv:1712.01815*, 2017. doi:https://doi.org/10.48550/arXiv.1712.01815. 30, 128

Slattery, D.; Amiel, S.; and Choudhary, P., Optimal Prandial Timing of Bolus Insulin in Diabetes Management: A Review, *Diabetic Medicine*, 35, 3, 306–316, Wiley Online Library, 2018. doi:10.1111/dme.13525. 2, 3, 14, 21

Sorensen, J. T., 1985. *A Physiologic Model of Glucose Metabolism in Man and its Use to Design and Assess Improved Insulin Therapies for Diabetes*. Ph.D. thesis, Massachusetts Institute of Technology. Available: https://dspace.mit.edu/handle/1721.1/15234. 41, 103, 105

Steil, G. M., Algorithms for a Closed-Loop Artificial Pancreas: The Case for Proportional-Integral-Derivative Control, *Journal of Diabetes Science and Technology*, 7, 6, 1621–1631, SAGE Publications, 2013. doi:10.1177/193229681300700623. 2, 18

Steil, G. M.; Palerm, C. C.; Kurtz, N.; Voskanyan, G.; Roy, A.; Paz, S.; and Kandeel, F. R., The Effect of Insulin Feedback on Closed Loop Glucose Control, *The Journal of Clinical Endocrinology and Metabolism*, 96, 5, 1402–1408, Oxford University Press, 2011. doi:10.1210/jc.2010-2578. 17, 18, 37

Steil, G. M.; Rebrin, K.; Darwin, C.; Hariri, F.; and Saad, M. F., Feasibility of Automating Insulin Delivery for the Treatment of Type 1 Diabetes, *Diabetes*, 55, 12, 3344–3350, American Diabetes Association, 2006. doi:10.2337/db06-0419. 17

Stenerson, M.; Cameron, F.; Wilson, D. M.; Harris, B.; Payne, S.; Bequette, B. W.; and Buckingham, B. A., The Impact of Accelerometer and Heart Rate Data on

Hypoglycemia Mitigation in Type 1 Diabetes, *Journal of Diabetes Science and Technology*, 8, 1, 64–69, SAGE Publications, 2014. doi:10.1177/1932296813516208. 100, 105, 106, 110, 142

Stewart, Z. A.; Wilinska, M. E.; Hartnell, S.; Temple, R. C.; Rayman, G.; Stanley, K. P.; Simmons, D.; Law, G. R.; Scott, E. M.; Hovorka, R.; et al., Closed-Loop Insulin Delivery During Pregnancy in Women with Type 1 Diabetes, *New England Journal of Medicine*, 375, 7, 644–654, Massachusetts Medical Society, 2016. doi:10.1056/NEJMoa1602494. 20, 21

Stoner, G. D., Hyperosmolar Hyperglycemic State, *American Family Physician*, 71, 9, 1723, 2005. Available: https://www.aafp.org/pubs/afp/issues/2017/1201/p729.html. 51, 83

Sun, Q.; Jankovic, M. V.; Budzinski, J.; Moore, B.; Diem, P.; Stettler, C.; and Mougiakakou, S. G., A Dual Mode Adaptive Basal-Bolus Advisor Based on Reinforcement Learning, *IEEE Journal of Biomedical and Health Informatics*, 23, 6, 2633–2641, IEEE, 2018. doi:10.1109/JBHI.2018.2887067. 34, 126

Sun, Q.; Jankovic, M. V.; and Mougiakakou, S. G., Reinforcement Learning-Based Adaptive Insulin Advisor for Individuals with Type 1 Diabetes Patients Under Multiple Daily Injections Therapy, *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3609–3612, IEEE, 2019. doi:10.1109/EMBC.2019.8857178. 34, 126

Sutton, R. S., Dyna, An Integrated Architecture for Learning, Planning, and Reacting, *ACM Sigart Bulletin*, 2, 4, 160–163, ACM, 1991. doi:https://doi.org/10.1145/122344.122377. 29, 33

Sutton, R. S., TD Models: Modeling the World at a Mixture of Time Scales. *Machine Learning Proceedings*, 531–539. Elsevier, 1995. doi:https://doi.org/10.1016/B978-1-55860-377-6.50072-4. 33, 63

Sutton, R. S. and Barto, A. G., 2018. *Reinforcement Learning: An Introduction*. MIT Press. 3, 4, 23, 28, 33, 46, 47, 68, 85

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; et al., DeepMind Control Suite, *arXiv:1801.00690*, 2018. doi:https://doi.org/10.48550/arXiv.1801.00690. 87, 88, 94

Teixeira, R. E. and Malin, S., The Next Generation of Artificial Pancreas Control Algorithms, *Journal of Diabetes Science and Technology*, 2, 1, 105–112, SAGE Publications, 2008. doi:10.1177/193229680800200115. 21, 95

Tejedor, M.; Woldaregay, A. Z.; and Godtliebsen, F., Reinforcement Learning Application in Diabetes Blood Glucose Control: A Systematic Review, *Artificial Intelligence in Medicine*, 104, 101836, Elsevier, 2020. doi:10.1016/j.artmed.2020.101836. 3, 23, 35

Tesauro, G., TD-Gammon, A Self-Teaching Backgammon Program, Achieves Master-Level Play, *Neural Computation*, 6, 2, 215–219, MIT Press, 1994. doi:10.1162/neco.1994.6.2.215. 29

Tessler, C.; Mankowitz, D. J.; and Mannor, S., Reward Constrained Policy Optimization, *arXiv:1805.11074*, 2018. doi:https://doi.org/10.48550/arXiv.1805.11074. 31

Teymourian, H.; Moonla, C.; Tehrani, F.; Vargas, E.; Aghavali, R.; Barfidokht, A.; Tangkuaram, T.; Mercier, P. P.; Dassau, E.; and Wang, J., Microneedle-Based Detection of Ketone Bodies Along With Glucose and Lactate: Toward Real-Time Continuous Interstitial Fluid Monitoring of Diabetic Ketosis and Ketoacidosis, *Analytical Chemistry*, 92, 2, 2291–2300, American Chemical Society, 2019. doi:https://doi.org/10.1021/acs.analchem.9b05109. 109

Thorndike, E., 1911. *Animal Intelligence: Experimental Studies.* Macmillan Press. doi:https://doi.org/10.5962/bhl.title.55072. 28

Tjoa, E. and Guan, C., A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI, *IEEE Transactions on Neural Networks and Learning Systems*, 32, 11, 4793–4813, IEEE, 2020. doi:10.1109/TNNLS.2020.3027314. 37

Todorov, E.; Erez, T.; and Tassa, Y., Mujoco: A Physics Engine for Model-Based Control, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033, IEEE, 2012. doi:10.1109/IROS.2012.6386109. 87, 88, 94

Topol, E., 2015. *The Patient Will See You Now: The Future of Medicine Is In Your Hands*. Basic Books. 130

Trief, P. M.; Cibula, D.; Rodriguez, E.; Akel, B.; and Weinstock, R. S., Incorrect Insulin Administration: A Problem That Warrants Attention, *Clinical Diabetes*, 34, 1, 25–33, American Diabetes Association, 2016. doi:10.2337/diaclin.34.1.25. 2, 4

Turchetta, M.; Berkenkamp, F.; and Krause, A., Safe Exploration In Finite Markov Decision Processes With Gaussian Processes, *Advances in Neural Information Processing Systems*, 29, 2016. Available: https://proceedings.neurips.cc/paper/2016/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf. 31

Turksoy, K.; Bayrak, E. S.; Quinn, L.; Littlejohn, E.; and Cinar, A., Adaptive Multivariable Closed-Loop Control of Blood Glucose Concentration in Patients With Type 1 Diabetes, *American Control Conference*, Proceedings (Pp. 2905–2910), IEEE, 2013a. doi:10.1109/ACC.2013.6580275. 100, 102, 103, 106, 107, 142

Turksoy, K.; Bayrak, E. S.; Quinn, L.; Littlejohn, E.; and Cinar, A., Multivariable Adaptive Closed-Loop Control of an Artificial Pancreas Without Meal and Activity Announcement, *Diabetes Technology and Therapeutics*, 15, 5, 386–400, 2013b. doi:10.1089/dia.2012.0283. 20, 25, 104, 107, 142

Turksoy, K.; Bayrak, E. S.; Quinn, L.; Littlejohn, E.; Rollins, D.; and Cinar, A., Hypoglycemia Early Alarm Systems Based On Multivariable Models, *Industrial and Engineering Chemistry Research*, 52, 35, 12329–12336, ACS Publications, 2013c. doi:10.1021/ie3034015. 24, 100, 103

Turksoy, K.; Hajizadeh, I.; Hobbs, N.; Kilkus, J.; Littlejohn, E.; Samadi, S.; Feng, J.; Sevil, M.; Lazaro, C.; Ritthaler, J.; Hibner, B.; Devine, N.; Quinn, L.; and Cinar, A., Multivariable Artificial Pancreas For Various Exercise Types and Intensities, *Diabetes Technology and Therapeutics*, 20, 10, 662–671, Mary Ann Liebert Inc., 2018. doi:10.1089/dia.2018.0072. 20, 103, 104, 107, 142

Turksoy, K.; Kilkus, J.; Hajizadeh, I.; Samadi, S.; Feng, J.; Sevil, M.; Lazaro, C.; Frantz, N.; Littlejohn, E.; and Cinar, A., Hypoglycemia Detection and Carbohydrate Suggestion in an Artificial Pancreas, *Journal of Diabetes Science and Technology*, 10, 6, 1236–1244, SAGE Publications, 2016. doi:10.1177/1932296816658666. 24, 103

Turksoy, K.; Quinn, L.; Littlejohn, E.; and Cinar, A., Multivariable Adaptive Identification and Control for Artificial Pancreas Systems, *IEEE Transactions on Biomedical Engineering*, 61, 3, 883–891, IEEE, 2013d. doi:10.1109/TBME.2013.2291777. 18, 19, 20, 100, 102, 106, 107, 142

Turksoy, K.; Quinn, L. T.; Littlejohn, E.; and Cinar, A., An Integrated Multivariable Artificial Pancreas Control System, *Journal of Diabetes Science and Technology*, 8, 3, 498–507, 2014. doi:10.1177/1932296814524862. 20, 104, 107, 142

Turksoy, K.; Samadi, S.; Feng, J.; Littlejohn, E.; Quinn, L.; and Cinar, A., Meal Detection in Patients With Type 1 Diabetes: A New Module for the Multivariable Adaptive Artificial Pancreas Control System, *IEEE Journal of Biomedical and Health Informatics*, 20, 1, 47–54, IEEE, 2015. doi:10.1109/JBHI.2015.2446413. 24, 103

Tyler, N. S. and Jacobs, P. G., Artificial Intelligence in Decision Support Systems for Type 1 Diabetes, *Sensors*, 20, 11, 3214, Multidisciplinary Digital Publishing Institute, 2020. doi:10.3390/s20113214. 24

US Food and Drug Administration, The content of Investigational Device Exemption (IDE) and Premarket Approval (PMA) Applications for Artificial Pancreas Device Systems, *Silver Spring*, 2012. 16

Vajapey, A., 2019. *Predicting Optimal Sedation Control With Reinforcement Learning*. Ph.D. thesis, Massachusetts Institute of Technology. Available: https://dspace.mit.edu/handle/1721.1/123126. 30, 94

vanRossum, G. and Drake, F. L., 2010. *Python Reference Manual*. Python Software Foundation: Amsterdam, Netherlands. 81

Vasiloglou, M. F.; Mougiakakou, S.; Aubry, E.; Bokelmann, A.; Fricker, R.; Gomes, F.; Guntermann, C.; Meyer, A.; Studerus, D.; and Stanga, Z., A Comparative Study on Carbohydrate Estimation: GoCARB vs. Dietitians, *Nutrients*, 10,

6, 741, Multidisciplinary Digital Publishing Institute, 2018. doi:10.3390/nu10060741. 24

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I., Attention is All You Need, *Advances in Neural Information Processing Systems*, 30, 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 120

Villena Gonzales, W.; Mobashsher, A. T.; and Abbosh, A., The Progress of Glucose Monitoring - A Review of Invasive to Minimally and Non-Invasive Techniques, Devices and Sensors, *Sensors*, 19, 4, 800, Multidisciplinary Digital Publishing Institute, 2019. doi:10.3390/s19040800. 13, 20, 21, 125

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al., Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning, *Nature*, 575, 7782, 350–354, Nature Publishing Group, 2019. doi:https://doi.org/10.1038/s41586-019-1724-z. 30

Visentin, R.; Campos-Náñez, E.; Schiavon, M.; Lv, D.; Vettoretti, M.; Breton, M.; Kovatchev, B. P.; Dalla Man, C.; and Cobelli, C., The UVA/Padova Type 1 Diabetes Simulator Goes From Single Meal to Single Day, *Journal of Diabetes Science and Technology*, 12, 2, 273–281, SAGE Publications, 2018. doi:10.1177/1932296818757747. 25, 41, 42

Vliebergh, J.; Lefever, E.; and Mathieu, C., Advances In Newer Basal and Bolus Insulins: Impact On Type 1 Diabetes, *Current Opinion in Endocrinology, Diabetes and Obesity*, 28, 1, 1–7, LWW, 2021. doi:10.1097/MED.0000000000000599. 13, 20, 21, 125

Wachi, A.; Sui, Y.; Yue, Y.; and Ono, M., Safe Exploration and Optimization of Constrained MDPs Using Gaussian Processes, *AAAI Conference on Artificial Intelligence*, Proceedings (Vol. 32, Pp. 1), 2018. doi:https://doi.org/10.1609/aaai.v32i1.12103. 31

Walsh, J.; Roberts, R.; and Bailey, T., Guidelines for Optimal Bolus Calculator Settings in Adults, *Journal of Diabetes Science and Technology*, 5, 1, 129–135, SAGE Publications, 2011. doi:10.1177/193229681100500118. 51

Weisman, A.; Bai, J. W.; Cardinez, M.; Kramer, C. K.; and Perkins, B. A., Effect of Artificial Pancreas Systems on Glycaemic Control in Patients with Type 1 Diabetes: A Systematic Review and Meta-Analysis of Outpatient Randomised Controlled Trials, *The Lancet Diabetes and Endocrinology*, 5, 7, 501–512, Lancet Publishing Group, 2017. doi:10.1016/S2213-8587(17)30167-5. 17

Weng, H.; Hettiarachchi, C.; Nolan, C.; Suominen, H.; and Lenskiy, A., Ensuring Security of Artificial Pancreas Device System Using Homomorphic Encryption, *Biomedical Signal Processing and Control*, 79, 104044, Elsevier, 2023. doi:https://doi.org/10.1016/j.bspc.2022.104044. 121

Weng, J.; Lin, M.; Huang, S.; Liu, B.; Makoviichuk, D.; Makoviychuk, V.; Liu, Z.; Song, Y.; Luo, T.; Jiang, Y.; et al., Envpool: A Highly Parallel Reinforcement Learning Environment Execution Engine, *arXiv:2206.10558*, 2022. doi:https://doi.org/10.48550/arXiv.2206.10558. 119

Wibral, M.; Vicente, R.; and Lindner, M., Transfer Entropy in Neuroscience, *Directed Information Measures in Neuroscience*, 3–36, Springer, 2014. doi:https://doi.org/10.1007/978-3-642-54474-3_1. 111

Wilinska, M. E.; Chassin, L. J.; Acerini, C. L.; Allen, J. M.; Dunger, D. B.; and Hovorka, R., Simulation Environment to Evaluate Closed-Loop Insulin Delivery Systems in Type 1 Diabetes, *Journal of Diabetes Science and Technology*, 4, 1, 132–144, SAGE Publications, 2010. doi:10.1177/193229681000400117. 25, 41

Wilinska, M. E.; Chassin, L. J.; Schaller, H. C.; Schaupp, L.; Pieber, T. R.; and Hovorka, R., Insulin kinetics in Type-1 Diabetes: Continuous and Bolus Delivery of Rapid Acting Insulin, *IEEE Transactions on Biomedical Engineering*, 52, 1, 3–12, IEEE, 2005. doi:10.1109/TBME.2004.839639. 103, 105

Woldaregay, A. Z.; Årsand, E.; Botsis, T.; Albers, D.; Mamykina, L.; and Hartvigsen, G., Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes, *Journal of Medical Internet Research*, 21, 5, e11030, JMIR Publications, 2019. doi:10.2196/11030. 3, 23

Wollstadt, P.; Lizier, J. T.; Vicente, R.; Finn, C.; Martinez-Zarzuela, M.; Mediano, P.; Novelli, L.; and Wibral, M., IDTxl: The Information Dynamics Toolkit xl: A Python Package for the Efficient Analysis of Multivariate Information Dynamics in Networks, *arXiv:1807.10459*, 2018. doi:https://doi.org/10.48550/arXiv.1807.10459. 111

Xiao, C.; Wu, Y.; Ma, C.; Schuurmans, D.; and Müller, M., Learning to Combat Compounding-Error in Model-Based Reinforcement Learning, *arXiv:1912.11206*, 2019. doi:https://doi.org/10.48550/arXiv.1912.11206. 28

Xie, J., Simglucose v0. 2.1 (2018) [online], 2018. Available: https://github.com/jxx123/simglucose. 34, 35, 42, 52, 117, 119

Xie, J. and Wang, Q., Benchmark Machine Learning Approaches with Classical Time Series Approaches on the Blood Glucose Level Prediction Challenge, *3rd International Workshop on Knowledge Discovery in Healthcare Data*, Proceedings (Pp. 97–102), 2018. Available: https://ceur-ws.org/Vol-2148/paper16.pdf. 25

Xie, J. and Wang, Q., Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type 1 Diabetes in Comparison with Classical Time-Series Models, *IEEE Transactions on Biomedical Engineering*, 67, 11, 3101–3124, IEEE, 2020. doi:10.1109/TBME.2020.2975959. 25

Xuegong, Z., Introduction to Statistical Learning Theory and Support Vector Machines, *Acta Automatica Sinica*, 26, 1, 32–42, 2000. 25

Yasini, S.; Naghibi-Sistani, M.; and Karimpour, A., Agent-Based Simulation for Blood Glucose Control in Diabetic Patients, *International Journal of Applied Science, Engineering and Technology*, 5, 1, 40–49, 2009. doi:doi.org/10.5281/zenodo.1060215. 34

Yassin, I. M.; Taib, M. N.; and Adnan, R., Recent Advancements & Methodologies In System Identification: A Review, *Scientific Research Journal*, 1, 1, 14–33, 2013. Available: http://www.scirj.org/papers-0813/scirj-august-2013-edition-03.pdf. 25

Yen, J. and Langari, R., 1999. *Fuzzy Logic: Intelligence, Control, and Information*, vol. 1. Prentice Hall. 19

Zaharieva, D. P.; Turksoy, K.; McGaugh, S. M.; Pooni, R.; Vienneau, T.; Ly, T.; and Riddell, M. C., Lag Time Remains With Newer Real-Time Continuous Glucose Monitoring Technology During Aerobic Exercise in Adults Living with Type 1 Diabetes, *Diabetes Technology and Therapeutics*, 21, 6, 313–321, Mary Ann Liebert, Inc., 2019. doi:10.1089/dia.2018.0364. 20, 110

Zahavy, T. et al., Learn What Not To Learn: Action Elimination With Deep Reinforcement Learning, *Advances in Neural Information Processing Systems*, 31, 2018. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/645098b086d2f9e1e0e939c27f9f2d6f-Paper.pdf. 31, 86, 94

Zakeri, I.; Adolph, A. L.; Puyau, M. R.; Vohra, F. A.; and Butte, N. F., Application Of Cross-Sectional Time Series Modeling for the Prediction of Energy Expenditure From Heart Rate and Accelerometry, *Journal of Applied Physiology*, 104, 6, 1665–1673, American Physiological Society, 2008. doi:10.1152/japplphysiol.01163.2007. 100

Zhu, T.; Li, K.; and Georgiou, P., A Dual-Hormone Closed-Loop Delivery System for Type 1 Diabetes Using Deep Reinforcement Learning, *arXiv:1910.04059*, 2019. doi:https://doi.org/10.48550/arXiv.1910.04059. 35, 40, 60, 87

Zhu, T.; Li, K.; Herrero, P.; and Georgiou, P., Basal Glucose Control in Type 1 Diabetes Using Deep Reinforcement Learning: An In Silico Validation, *IEEE Journal of Biomedical and Health Informatics*, 25, 4, 1223–1232, IEEE, 2020. doi:10.1109/JBHI.2020.3014556. 35, 40, 60, 87