MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI250997

Comparing Deterministic and Stochastic Reinforcement Learning for Glucose Regulation in Type 1 Diabetes

David TIMMS[a], Chirath HETTIARACHCHI[a,1], and Hanna SUOMINEN[a,b]
[a] The Australian National University, Australia
[b] University of Turku, Finland
ORCiD ID: DT https://orcid.org/0009-0000-6683-5914, CH https://orcid.org/0000-0002-7702-0718, HS https://orcid.org/0000-0002-4195-1641

Abstract. Type 1 Diabetes (T1D) is a chronic condition affecting millions worldwide, requiring external insulin administration to regulate blood glucose levels and prevent serious complications. Artificial Pancreas Systems (APS) for managing T1D currently rely on manual input, which adds a cognitive burden on people with T1D and their carers. Research into alleviating this burden through Reinforcement Learning (RL) explores enabling the APS to autonomously learn and adapt to the complex dynamics of blood glucose regulation, demonstrating improvements in *in-silico* evaluations compared to traditional clinical approaches. This evaluation study compared the primary polarities of RL for glucose regulation, namely, stochastic (e.g., Proximal Policy Optimization (PPO) and deterministic (e.g., Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithms *in-silico* using quantitative and qualitative methods, patient specific clinical metrics, and the adult and adolescent cohorts of the U.S. Food and Drug Administration approved UVA/PADOVA 2008 model. Although the behavior of TD3 was easier to interpret, it did not typically outperform PPO, thereby challenging assessing their safety and suitability. This conclusion highlights the importance of improving RL algorithms in APS applications for both interpretability and predictive performance in future research.

Keywords. Artificial Pancreas, Deep Learning, Evaluation Study, Type 1 Diabetes

1. Introduction

Type 1 Diabetes (T1D) is a disease that affects millions of people worldwide, limiting their ability to regulate blood glucose levels [1]. People with T1D are dependent on external insulin administration to maintain blood glucose levels within the optimal normoglycemic range (70-180 mg/dL) [2]. Failure to regulate blood glucose can lead to conditions such as cardiovascular or renal disease, blindness, limb amputations, and, in extreme cases, even death [2]. Artificial Pancreas Systems (APS) contribute to latest treatment approaches, referring to an external wearable device that includes both a sensor to continuously measure blood glucose levels and a pump to continuously infuse the required insulin dose, based on a control algorithm, which calculates this dose subject to keeping blood glucose levels in the optimal range. Current commercialized APSs deploy Proportional Integral Derivative (PID) control, Model Predictive Control (MPC), and other classical control methods, and as such, to operate effectively, they require

[1] **Corresponding Author:** Dr Chirath Hettiarachchi, The Australian National University (ANU), 145 Science Rd, Acton ACT 2601, Australia, chirath.hettiarachchi@anu.edu.au.

burdensome manual input of, e.g., planned exercise and carbohydrate intake from the person with T1D and their careers, but research into fully automatic APSs are emerging to improve the lives of people with T1D and their carers [3-5].

Reinforcement Learning (RL) studies are enabling APSs to fully automatically learn a decision-making strategy for insulin dosing with their *in-silico* evaluations demonstrating improvements over classical control algorithms and traditional clinical approaches [3-8]. RL algorithms belong to either Machine or Deep Learning (M/DL) and can be broadly categorized as either stochastic or deterministic: The insulin-dosing strategy learned by stochastic algorithms is subject to some randomness, e.g., by employing stochastic policy gradient RL algorithms, including Proximal Policy Optimization (PPO), Soft Actor Critic (SAC), and Glucose Control by Glucose Prediction and Planning (G2P2C) [3-5, 8]. The behavior of deterministic RL is more predictable with recent works highlighting Deterministic Policy Gradient (DPG), Deep Deterministic Policy Gradient (DDPG), and Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithms [6, 7].

We compare stochastic and deterministic RL algorithms for glucose regulation in T1D, also benchmarking them against a clinical treatment approach (i.e., Basal-Bolus Ideal, BBI) [8]. We evaluate them *in-silico* based on underlying characteristics, predictive performance, and clinical applicability using the adult and adolescent cohorts of a US Food and Drug Administration (FDA) approved model [9, 10].

2. Methods

Following the best practice for developing control algorithms (because APSs are high-risk medical devices, FDA-approved simulators are deployed to conduct typical daily simulations (incl. meal protocols) on *in-silico* T1D subjects wearing commercially available glucose sensors and pumps to predict clinical outcomes and inform clinical trial designs), we used the *in-silico* adult and adolescent cohorts of the Simglucose simulator [9] which is based on the FDA-approved UVA/PADOVA 2008 model [10].

Following the experiment setup of previous studies [5, 8] (see these papers for a detailed problem formulation and implementation), we chose for this evaluation study the Insulet insulin pump and the GuardianRT glucose sensor, operating with a 5-minute sampling interval, and implemented a standardized meal protocol to simulate realistic meal scenarios, increasing the complexity and variability of the RL task to regulate glucose. The evaluation meal protocol spanned 24 hours from 00:00, fixed with three meals: 40g of CHO for breakfast at 8:00, 80g of CHO for lunch at 13:00, and 60g of CHO for dinner at 20:00. The training protocol involved randomizing mealtimes and carbohydrate content to introduce uncertainty and challenge the RL algorithms. Also based on [5, 8], we selected the best performing stochastic and deterministic RL algorithms (i.e., PPO and TD3, respectively) for this evaluation study. For further clinical treatment context, we included a Basal-Bolus insulin infusion approach in the evaluation and implemented a version of it as our BBI algorithm [8]. Unlike PPO and TD3, it required perfect manual meal announcements and carbohydrate estimations without considering any human error and thus formed our gold standard benchmark.

Our evaluations used quantitative (clinical and statistical performance metrics) and qualitative (interpretability and clinical analyses) methods for 20 *in-silico* patients (10 adolescents and 10 adults). We ran 1,500 evaluation simulations for each subject to address different patient characteristics and initial glucose levels. Our primary clinical

objective in glucose regulation was to increase the Time spent In the normoglycemic Range (TIR) while minimizing catastrophic failures, defined as glucose levels outside the detectable range (40-600 mg/dL) of the sensor. We calculated Failure Rate (FR) as a percentage of such occurrences. Finally, we examined the interpretability of PPO and TD3 by exploring the correlation patterns between simulated I_t, defined as the insulin dose at time t, and glucose and insulin values of the previous 1-hour window (I_{t-1}, I_{t-2}, ... , I_{t-12}, G_{t-1}, G_{t-2}, ... , G_{t-12}), which served also as the input feature vector in RL. Our goal was to investigate whether RL was effective in learning a meaningful relationship between the historical feature vector to use past information to administer insulin, anticipating changes in blood glucose levels, ensuring stable regulation.

3. Results

PPO and TD3 did not tend to outperform the clinical benchmark BBI in the TIR and FR metrics, with PPO usually outperforming TD3 (**Table 1**). However, BBI required perfect meal announcements and carbohydrate estimates, which were not inputs to RL. We anticipated that their inclusion also in RL would improve the performance of PPO and TD3, but this would contradict fully automating APSs.

Contrary to the fairly random and as such, hard to interpret, insulin response curves for the stochastic PPO algorithm, those for the deterministic TD3 algorithm were smooth, with each point closely related to the previous one (**Fig. 1**). As expected, insulin administered by PPO had negligible linear correlation (±0.3) with either glucose or insulin levels over the previous one-hour period **(Fig. 2)**. In comparison, TD3 showed strong positive correlation with past insulin and glucose.

4. Discussion

Stochastic PPO performs the best in this study, but its insulin response curve shows significant randomness, making its behavior hard to interpret. This randomness is inherent to stochastic RL methods, which sample actions (e.g., insulin) from a probability distribution. In contrast, deterministic TD3 exhibits a smooth administered insulin curve and is much more predictable. While this smoothness ensures stability, it limits the adaptability of this method, reducing its ability to quickly adjust insulin delivery in response to sudden changes in blood glucose, which can challenge effective regulation. This is evident through the better performance of PPO compared to TD3.

Table 1. Performance comparison of PPO, TD3, and BBI algorithms

Algorithm	Adolescents (TIR / FR)	Adults (TIR / FR)
BBI	71.43 ± 12.31 / 0%	71.02 ± 11.29 / 0.39%
PPO	63.72 ± 13.95 / 4.93%	69.12 ± 10.53 / 2.79%
TD3	60.99 ± 19.18 / 25.6%	62.79 ± 16.30 / 18.37%

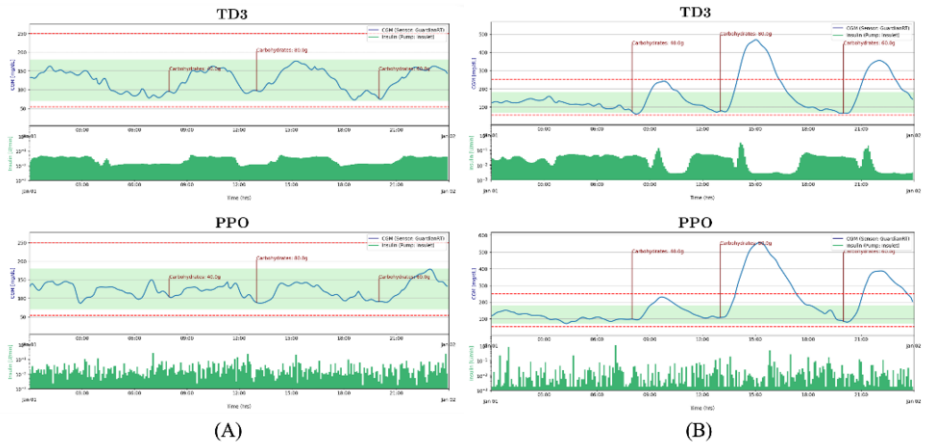

Figure 1. Comparing simulations with TD3 and PPO algorithms for (A) Adolescent0 and (B) Adolescent6. In each 1-day simulation insulin (bottom) is administered to the body based on the RL algorithm to control glucose levels (top). Ideally, they should stay within the normoglycemic range (green shading). Vertical lines indicate the timing and carbohydrate content of meals. Graphs for all 20 *in-silico* patients, with a larger font size for readability, can be accessed at our publicly available tool at https://capsml.com/.

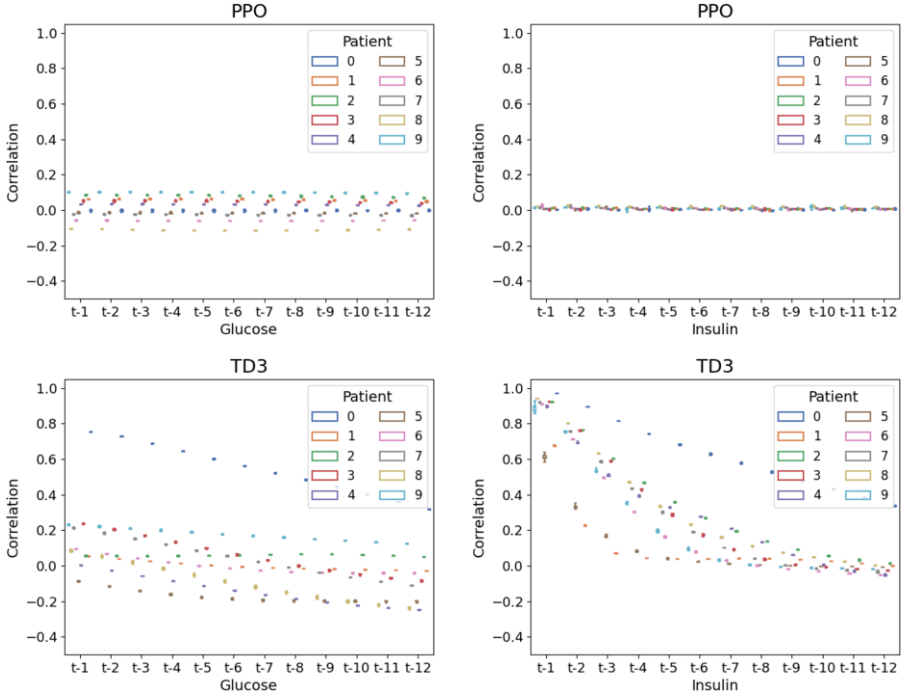

Figure 2. Correlation analysis of currently administered insulin dose and past glucose and insulin values (previous hour) for the adolescent cohort (10 subjects) and two RL algorithms (PPO and TD3)

In APSs, due to the safety critical nature, more predictable and interpretable methods are preferred [8]. For TD3, our study has identified that more recent blood glucose and insulin data have a higher correlation with administered insulin, aligning with the human body's natural response dynamics to insulin [2]. Such insights would be valuable for interpreting and understanding RL, or more generally, M/DL in clinical applications.

This study was limited to analyzing the characteristics of stochastic and deterministic RL algorithms for glucose regulation; a comprehensive comparison of the proposed RL algorithms with prior work was conducted in our previous study [8].

5. Conclusion

Although the behavior of TD3 is easier to interpret, this RL algorithm does not always outperform PPO. This conclusion challenges assessing algorithmic safety and suitability, also highlighting the importance of improving APS applications for both interpretability and predictive performance in future research. Because this study was restricted to the *in-silico* adult and adolescent cohorts, future research could focus on extending the analysis to the challenging child cohort.

Acknowledgment

This research was delivered in partnership with Our Health in Our Hands, a strategic initiative of ANU, which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians, and health-care providers. We gratefully acknowledge funding from the MRFF 2022 National Critical Research Infrastructure (MRFCRI000138, Developing a new digital therapeutic or depression: Closed loop non-invasive brain stimulation). This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure under the ANU Merit Allocation Scheme (ny83 and eu59) and ANU Startup Scheme (sj53).

References

[1]. Gregory, Gabriel A., 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) (2022): 741-760.
[2]. DiMeglio, Linda A., Carmella Evans-Molina, and Richard A. Oram. "Type 1 diabetes." *The Lancet* 391(10138) (2018): 2449-2462.
[3]. Fox, Ian, et al. "Deep reinforcement learning for closed-loop blood glucose control." *Proceedings of Machine Learning Research* 126 (2020):1-28.
[4]. Lee, Seunghyun, et al. "Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation." *IEEE Journal of Biomedical & Health Informatics* 25(2) (2020): 536-546.
[5]. Hettiarachchi, Chirath, et al. "A reinforcement learning based system for blood glucose control without carbohydrate estimation in type 1 diabetes: In silico validation." *Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society* (2022): 950-956.
[6]. Emerson, Harry, Matthew Guy, and Ryan McConville. "Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes." *Journal of Biomedical Informatics* 142 (2023): 104376.
[7]. Zhu, Taiyu, Kezhi Li, and Pantelis Georgiou. "Offline deep reinforcement learning and off-policy evaluation for personalized basal insulin control in type 1 diabetes." *IEEE Journal of Biomedical & Health Informatics* 27(10) (2023): 5087-5098.
[8]. Hettiarachchi, Chirath, et al. "G2P2C—A modular reinforcement learning algorithm for glucose control by glucose prediction and planning in Type 1 Diabetes." *Biomedical Signal Processing & Control* 90 (2024): 105839.
[9]. Xie, Jinyu, 2018. Simglucose v0.2.1. https://github.com/jxx123/simglucose.
[10]. Kovatchev, Boris P., et al. "In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes." *Journal of Diabetes Science & Technology* 3(1) (2009): 44-55.