ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Reinforcement Learning for Personalized Insulin Dosing: A Comparative Study of A2C, SAC and PPO on Real-World Clinical Data

Chinatu M. Anyanwu[1], Nkiru C. Ogbonna[2*], Mary Ofuru Kam[3], Stephen Uche Udeh[4], Ogechi Gift Onyedi[5]

[1]Faculty of Computing, Maduka University, Ekwegbe-Enugu State

[2]Department of ICT/Innovation Centre, University of Nigeria, Nsukka

[3]Department of Computer Science, Veritas University, Bwari, Abuja, Nigeria

[4]Department of Computer Science, University of Nigeria, Nsukka

[5]School of Health Science, Maduka University, Ekwegbe-Enugu State

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150700087

Received: 16 July 2026; Accepted: 21 July 2026; Published: 13 August 2026

ABSTRACT

Personalized insulin dosing for Type 1 diabetes mellitus (T1DM) remains challenging because of complex glucose–insulin dynamics and substantial patient variability. Reinforcement learning (RL) has emerged as a promising approach for adaptive insulin management, yet the reliability of learned policies depends heavily on reward design and evaluation strategy. This study compares three actor–critic RL algorithms: Soft Actor-Critic (SAC), Advantage Actor-Critic (A2C), and Proximal Policy Optimization (PPO) for personalized insulin dosing using real-world continuous glucose monitoring, insulin delivery, basal insulin, and meal intake data from the OhioT1DM dataset. A custom Gymnasium-based environment was developed, and all algorithms were trained under identical conditions for 100,000 timesteps. Performance was evaluated using cumulative reward together with clinically relevant measures, including Time in Range (TIR) and insulin dosing behaviour. Although A2C and PPO achieved higher cumulative rewards than SAC, both converged to near-zero insulin dosing policies that exploited the reward formulation rather than learning clinically meaningful glucose regulation. In contrast, SAC maintained adaptive dosing behaviour, achieving a TIR of 72.71% with an average insulin dose of 1.769 U/step. These findings show that higher cumulative reward does not necessarily correspond to better clinical decision-making in open-loop reinforcement learning environments. The study highlights the importance of behaviour-focused evaluation alongside conventional reward metrics and provides practical insights for developing safer and more reliable reinforcement learning systems for personalized diabetes management.

Keywords: Reinforcement Learning, A2C, SAC, PPO, Insulin dosing, Type 1 Diabetes, OhioT1DM Dataset, Reward Hacking, Time in Range, Clinical Decision Support.

INTRODUCTION

Type 1 diabetes mellitus (T1DM) requires continuous monitoring of blood glucose levels and timely insulin administration to maintain glucose within a safe therapeutic range. Despite advances in continuous glucose monitoring (CGM) and automated insulin delivery technologies, achieving optimal glycaemic control remains difficult because glucose responses vary considerably among individuals and are influenced by factors such as meals, physical activity, stress, and daily physiological changes. These challenges highlight the need for decision-support systems capable of adapting insulin therapy to individual patient needs.

Reinforcement learning (RL) has emerged as a promising approach for personalized healthcare decision-making because it enables an agent to learn sequential actions through interactions with an environment. Unlike supervised learning, which predicts outcomes from labelled data, RL continuously learns actions that maximize

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

long-term objectives, making it well suited to insulin dosing where treatment decisions have delayed and cumulative effects.

Recent studies have demonstrated the potential of RL for automated insulin delivery using both physiological simulators and real-world datasets. However, most existing work focuses primarily on improving glucose control and cumulative reward, with limited attention to whether the learned policies represent clinically meaningful behaviour. In particular, open-loop environments based on historical patient data introduce an important challenge because an agent's actions do not influence subsequent physiological states. Under these conditions, a policy may maximize reward by exploiting weaknesses in the reward function rather than learning an effective insulin dosing strategy.

To investigate this challenge, this study compares three widely used actor-critic reinforcement-learning algorithms: Soft Actor-Critic (SAC), Advantage Actor-Critic (A2C), and Proximal Policy Optimization (PPO) for personalized insulin dosing using real-world data from the OhioT1DM dataset. A reproducible Gymnasium-based environment was developed, and algorithm performance was evaluated using both cumulative reward and clinically relevant behavioural measures.

The main contributions of this study are:

i. A comparative evaluation of SAC, A2C, and PPO for personalized insulin dosing using real-world clinical data under identical experimental conditions.
ii. An empirical investigation of reward hacking in an open-loop clinical reinforcement learning environment, demonstrating how reward optimization can produce clinically undesirable dosing behaviour.
iii. A behaviour-focused evaluation framework that combines reward-based performance with clinical metrics and policy behaviour to provide a more comprehensive assessment of reinforcement learning agents.
iv. Practical insights into reward design and evaluation strategies for developing safer reinforcement learning systems for personalized diabetes management.

By demonstrating how reward optimization can diverge from clinically meaningful decision-making, this study contributes to the development of more reliable reinforcement learning methods for healthcare decision support.

Related Work

Reinforcement learning (RL) has become an important approach for personalized healthcare because it enables sequential decision-making that can adapt to changing patient conditions. Unlike conventional machine learning methods that focus primarily on prediction, RL learns treatment policies through interactions with an environment, making it particularly suitable for insulin dosing where decisions have cumulative and time-dependent effects. Similarly, personalized glucose prediction models have demonstrated that individualized approaches can improve glycaemic management compared with generalized models (Parveen, 2021).

Several actor-critic algorithms have been developed for continuous control problems. Mnih et al. (2016) introduced the Advantage Actor-Critic (A2C) framework, while Schulman et al. (2017) proposed Proximal Policy Optimization (PPO), which improves training stability through a clipped policy update mechanism. Haarnoja et al. (2018) later introduced Soft Actor-Critic (SAC), an off-policy algorithm that incorporates entropy regularization to improve exploration and policy robustness. Beyond these foundational methods, Zhao et al. (2019) demonstrated that asynchronous deep reinforcement learning with dynamic weight updating can improve learning efficiency in complex environments, while Zheng et al. (2025) proposed a reward-based prioritization strategy for PPO that further enhances policy optimization. These developments illustrate the continued evolution of reinforcement learning algorithms for challenging sequential decision-making tasks.

Recent studies have applied reinforcement learning to automated insulin delivery and glucose regulation with encouraging results. Singh and Raj (2023) developed a deep learning-based artificial pancreas using real-world patient data to improve glycaemic control. Lei et al. (2024) proposed an adaptive glucose control framework that incorporated temporal dependencies to support personalized insulin management, while Dénes-Fazakas et al. (2024) demonstrated clinically relevant Time in Range (TIR) performance using a PPO-based glucose control

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

strategy. Collectively, these studies highlight the potential of reinforcement learning for diabetes management and intelligent clinical decision support.

Despite these advances, important limitations remain. Many existing studies rely on physiological simulators that simplify glucose-insulin dynamics and may not fully capture the variability observed in real-world clinical data. Furthermore, evaluation is often centred on cumulative reward or glucose outcome metrics, providing limited insight into whether the learned policies reflect clinically meaningful insulin dosing behaviour. As a result, policies that achieve favourable numerical performance may still exploit weaknesses in the reward formulation rather than learning appropriate treatment strategies.

This study addresses these limitations by comparing SAC, A2C, and PPO using real-world OhioT1DM data within a reproducible open-loop reinforcement learning environment. Rather than evaluating algorithms solely on cumulative reward or glucose outcomes, the study examines policy behaviour alongside clinical performance to determine whether the learned dosing strategies are clinically meaningful. By explicitly investigating reward hacking in an open-loop clinical setting, this work provides new evidence on the importance of reward design and behaviour-focused evaluation for developing safer reinforcement learning systems in personalized diabetes management.

METHODOLOGY

Dataset and Preprocessing

This study utilized the OhioT1DM dataset, a publicly available real-world dataset containing continuous glucose monitoring (CGM), insulin delivery, physiological measurements, and lifestyle information collected from individuals with Type 1 diabetes mellitus (T1DM). The dataset has been widely adopted in diabetes research because it provides high-resolution longitudinal data that capture real patient glucose management patterns.

To ensure a controlled and reproducible experimental setting, data from a single patient over a continuous one-month period were selected. Four data streams were extracted for analysis: continuous glucose measurements, bolus insulin administration, basal insulin delivery, and meal intake records. These streams were synchronized and resampled into uniform five-minute intervals to create a sequential decision-making environment suitable for reinforcement learning.

Table 1: Summary of OhioT1DM Dataset features used in the Study

Data Component	Description	Sampling Interval	Role in RL Environment
Continuous Glucose Monitoring (CGM)	Blood glucose measurements from the patient	5 minutes	Primary physiological state indicator
Bolus Insulin	Patient-administered insulin doses	5 minutes	Historical insulin administration information
Basal Insulin Rate	Continuous background insulin delivery	5 minutes	Represents baseline insulin requirements
Meal Intake	Recorded meal information affecting glucose variation	5 minutes	External factor influencing glucose dynamics

Missing CGM measurements were handled using forward filling based on the most recent observation, while glucose values were constrained to a physiologically plausible range of 40–400 mg/dL to minimize the influence

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

of sensor artefacts and outliers. After preprocessing, the dataset comprised 8,928 sequential time steps with a mean glucose level of 146.48 mg/dL (SD = 57.20), a mean basal insulin rate of 0.36 U/h, and a mean meal intake of 2.72 kcal per five-minute interval. Because the dataset consists of historical patient observations, the resulting reinforcement learning environment is open-loop, meaning that the agent's insulin actions do not influence subsequent glucose measurements.

Environment Design

A custom Gymnasium environment was developed to model personalized insulin dosing as a sequential decision-making problem. At each decision step, the agent observes the patient's current physiological state and selects an insulin dose with the objective of maintaining blood glucose within the recommended clinical range.

The observation space consists of five normalized features: three consecutive continuous glucose monitoring (CGM) readings, the current basal insulin rate, and meal intake. Including multiple glucose measurements enables the agent to capture short-term glucose trends rather than relying on a single observation.

The action space is continuous, allowing the agent to generate values between -1 and 1, which are linearly mapped to insulin doses ranging from 0 to 10 units every five minutes. This formulation provides sufficient flexibility to learn individualized dosing strategies.

The reward function encourages clinically desirable glucose regulation by assigning a reward of +1 when glucose remains within the target range (70–180 mg/dL), while hypoglycaemic (<70 mg/dL) and hyperglycaemic (>180 mg/dL) events receive penalties of -3 and -1, respectively. An additional dose-dependent penalty (-0.1 × insulin dose) discourages unnecessary insulin administration. Because the environment is open-loop, insulin actions do not alter subsequent glucose measurements. Consequently, minimizing insulin administration can increase cumulative reward without improving glucose regulation, creating the conditions under which reward hacking may occur.

Each training episode begins from a randomly selected point within the dataset to expose the agent to diverse glucose patterns. Episodes continue for a maximum of 288 steps, representing 24 hours of patient data, and terminate earlier if the end of the dataset is reached or glucose values fall outside predefined safety limits.

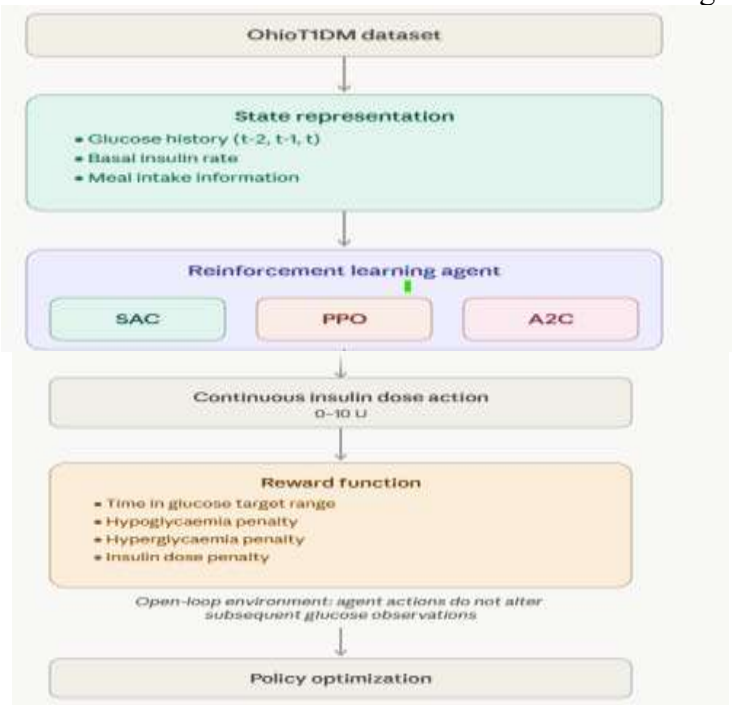

Figure 1: Overview of the proposed reinforcement learning framework for personalized insulin dosing, showing the interaction between the open-loop environment, the RL agent (A2C, SAC, or PPO), and the reward function.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Algorithms and Training

Three actor-critic reinforcement learning algorithms were evaluated in this study: Advantage Actor-Critic (A2C), Proximal Policy Optimization (PPO), and Soft Actor-Critic (SAC). A2C and PPO represent on-policy learning approaches, whereas SAC is an off-policy algorithm that incorporates entropy regularization to encourage exploration. All algorithms were implemented using the Stable-Baselines3 library and trained within the same custom Gymnasium environment to ensure a fair comparison. The environment was vectorized using DummyVecEnv, and each agent was trained for 100,000 timesteps using the default Stable-Baselines3 hyperparameters for its respective algorithm. No algorithm-specific hyperparameter tuning was performed so that differences in performance could be attributed primarily to the learning algorithms rather than parameter optimization. We acknowledge this as a limitation of the current study because default hyperparameters were used uniformly. It remains possible that tuned variants of A2C or PPO could partially mitigate the zero-dosing convergence reported in section 5, and this possibility is addressed further in section 6.

During training, episode reward, average glucose level, and average insulin dose were recorded to monitor learning behaviour and policy convergence. Using identical training conditions for all three algorithms enabled a consistent comparison of both reward optimization and learned insulin dosing behaviour.

Experimental Setup

All experiments were implemented in Python using Gymnasium to develop the reinforcement learning environment and Stable-Baselines3 to implement the A2C, PPO, and SAC algorithms. Data preprocessing and analysis were performed using NumPy and pandas, while Matplotlib was used for visualization.

Training Configuration

Each algorithm was trained under identical experimental conditions to ensure a fair comparison. Training was conducted for 100,000 timesteps in the same environment using a maximum episode length of 288 steps, corresponding to 24 hours of patient data. The observation space comprised five normalized features representing recent glucose history, basal insulin rate, and meal intake, while the continuous action space generated insulin doses between 0 and 10 units. Episodes terminated either after reaching the maximum episode length or when glucose values exceeded predefined safety limits.

Table 2: Training Configuration used for all reinforcement learning algorithms

Parameter	Value
Total training timesteps (per algorithm)	100,000
Episode length (max)	288 steps (24 h)
Observation space	5-dimensional (3 × glucose history, basal rate, meal intake)
Action space	Continuous, [-1, 1] → mapped to [0, 10] U
Vectorization	DummyVecEnv (single environment)
Reward range per step	-3.0 to +1.0, plus -0.1 × dose penalty
Unsafe termination bounds	< 60 mg/dL or > 300 mg/dL
Algorithms compared	A2C, PPO, SAC (Stable-Baselines3 defaults)

Evaluation Protocol

Following training, each policy was evaluated over ten independent episodes using deterministic action selection to assess learned behaviour without exploration noise. Each evaluation episode was limited to 100 steps, and cumulative reward, average glucose level, and average insulin dose were recorded.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

To assess clinical relevance, additional performance measures including Time in Range (TIR), hypoglycaemia rate, hyperglycaemia rate, mean glucose deviation, and average insulin dose were calculated. Combining reward-based and clinically meaningful metrics enabled a comprehensive assessment of policy performance and helped identify cases where high cumulative reward did not correspond to clinically appropriate insulin dosing

.Table 3: Evaluation Metrics Used in the Study

Metric	**Definition**	**Clinical Relevance**
Cumulative Reward	Total reward accumulated during evaluation	Overall policy optimization
Time in Range (TIR)	Percentage of glucose values between 70 and 180 mg/dL	Glycaemic control
Hypoglycaemia Rate	Percentage of glucose values below 70 mg/dL	Safety assessment
Hyperglycaemia Rate	Percentage of glucose values above 180 mg/dL	Hyperglycaemia assessment
Mean Glucose Deviation	Average deviation from the target glucose range	Glucose regulation quality
Average Insulin Dose	Mean insulin administered per step	Learned dosing behaviour

RESULTS AND DISCUSSION

Training Performance Analysis

The learning behaviour of SAC, A2C, and PPO over the 100,000 training timesteps is illustrated in Figure 2. Although all three algorithms converged to comparable reward levels, their learned insulin dosing strategies differed substantially.

SAC maintained an adaptive dosing policy throughout training, with insulin administration stabilizing at approximately 1–2 U per step after an initial exploration phase. In contrast, both A2C and PPO progressively reduced insulin administration, ultimately converging to near-zero dosing policies.

Despite their different optimization mechanisms, the two on-policy algorithms exhibited remarkably similar learning behaviour, suggesting that both exploited the reward structure rather than learning clinically meaningful insulin regulation.

Average glucose levels remained similar across all three algorithms because the environment replayed historical patient data, meaning that insulin actions did not influence subsequent glucose observations. Consequently, the observed differences in performance were driven primarily by the learned dosing policies rather than changes in glucose trajectories.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

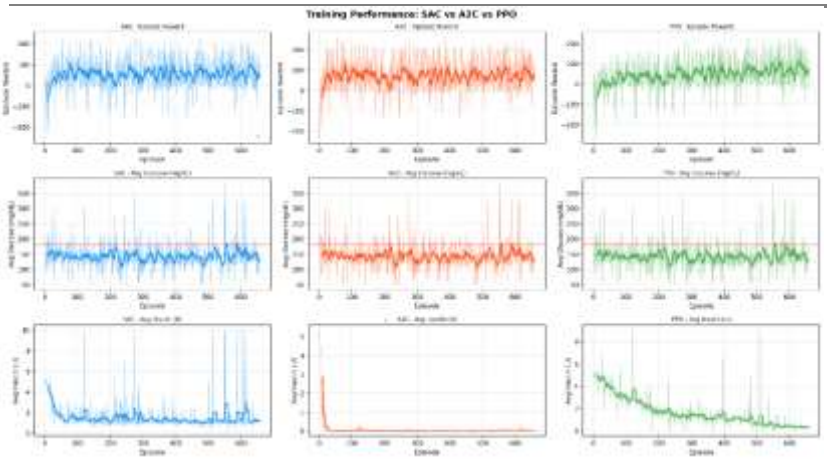

Figure 2: Training performance of SAC, PPO, and A2C over 100,000 training timesteps showing episode reward, average glucose level, and average insulin dose.

Reward Comparison

Table 4 summarizes the average evaluation performance of the three reinforcement learning algorithms across ten independent evaluation episodes. Although A2C and PPO achieved the highest mean cumulative reward (53.60), both converged to near-zero insulin dosing policies. In contrast, SAC obtained a lower mean reward (43.03) while maintaining an average insulin dose of 1.769 U per step.

Table 4: Mean evaluation performance of SAC, A2C, and PPO across ten evaluation episodes.

Metric	SAC	A2C	PPO
Mean total reward (10 episodes)	43.03	53.60	53.60
Mean glucose (mg/dL)	146.5	146.5	146.5
Mean insulin dose (U/step)	1.769	≈0.00	≈0.00

The relationship between cumulative reward, average glucose level, and insulin dosing behaviour is illustrated in Figure 3. The reward curves show that A2C and PPO consistently outperformed SAC in terms of cumulative reward, whereas average glucose remained almost identical across all three algorithms. The principal difference was observed in insulin administration, where SAC continued to produce adaptive dosing decisions while A2C and PPO generated almost no insulin throughout the evaluation period.

These results indicate that reward-based performance differed substantially from the learned insulin dosing behaviour.

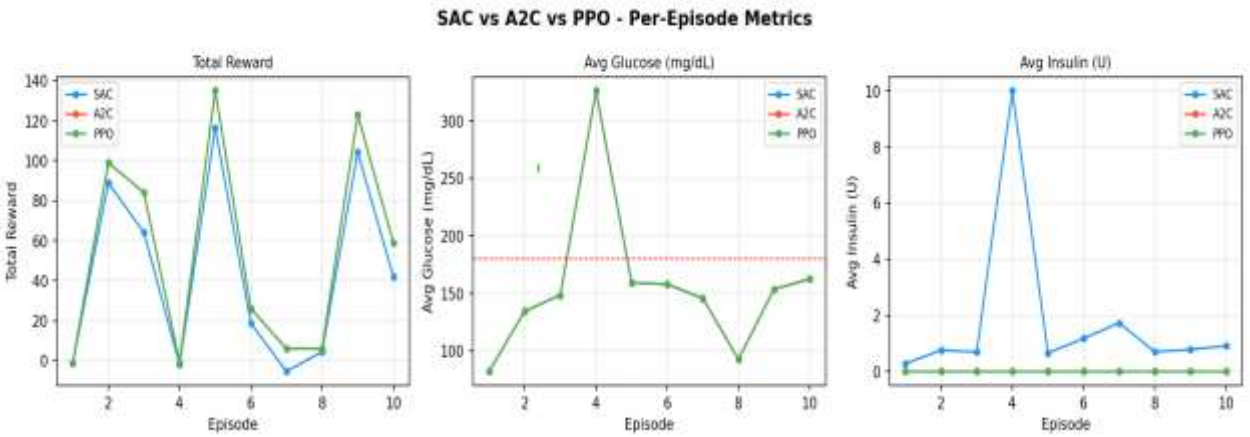

Figure 3: Per-episode total reward, average glucose, and average insulin dose for SAC vs. A2C across 10 evaluation episodes.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

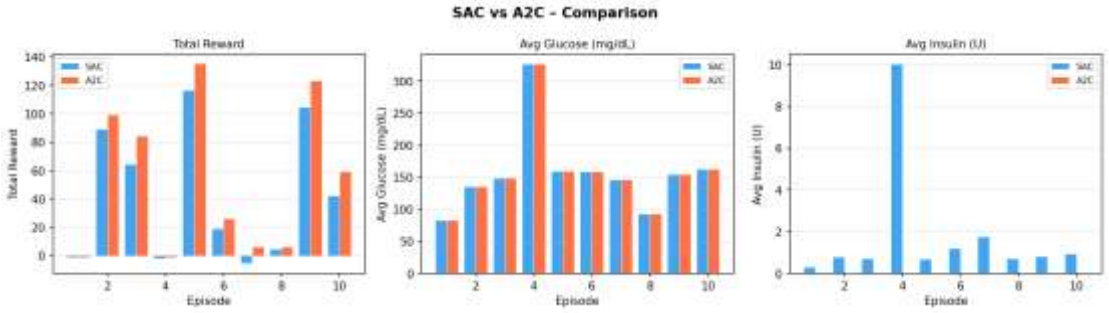

Figure 4: Comparison of cumulative reward, average glucose, and average insulin dose across the three reinforcement learning algorithms.

Evaluation Trajectories

Figure 5 presents representative glucose and insulin dose trajectories during policy evaluation. As expected in an open-loop environment, the glucose trajectories were identical for all three algorithms because the recorded patient data were replayed irrespective of the agent's actions. In contrast, the insulin dosing trajectories revealed clear behavioural differences.

SAC generated adaptive insulin doses that varied in response to the observed glucose profile, indicating that the learned policy actively responded to changes in the patient's physiological state. Conversely, both A2C and PPO produced near-zero insulin doses throughout the evaluation period, regardless of changes in glucose levels. The similarity of these two policies suggests that both on-policy algorithms converged to the same low-dosing strategy despite differences in their optimization procedures.

Examining the evaluation trajectories provides additional insight into the behavioural differences among the three algorithms.

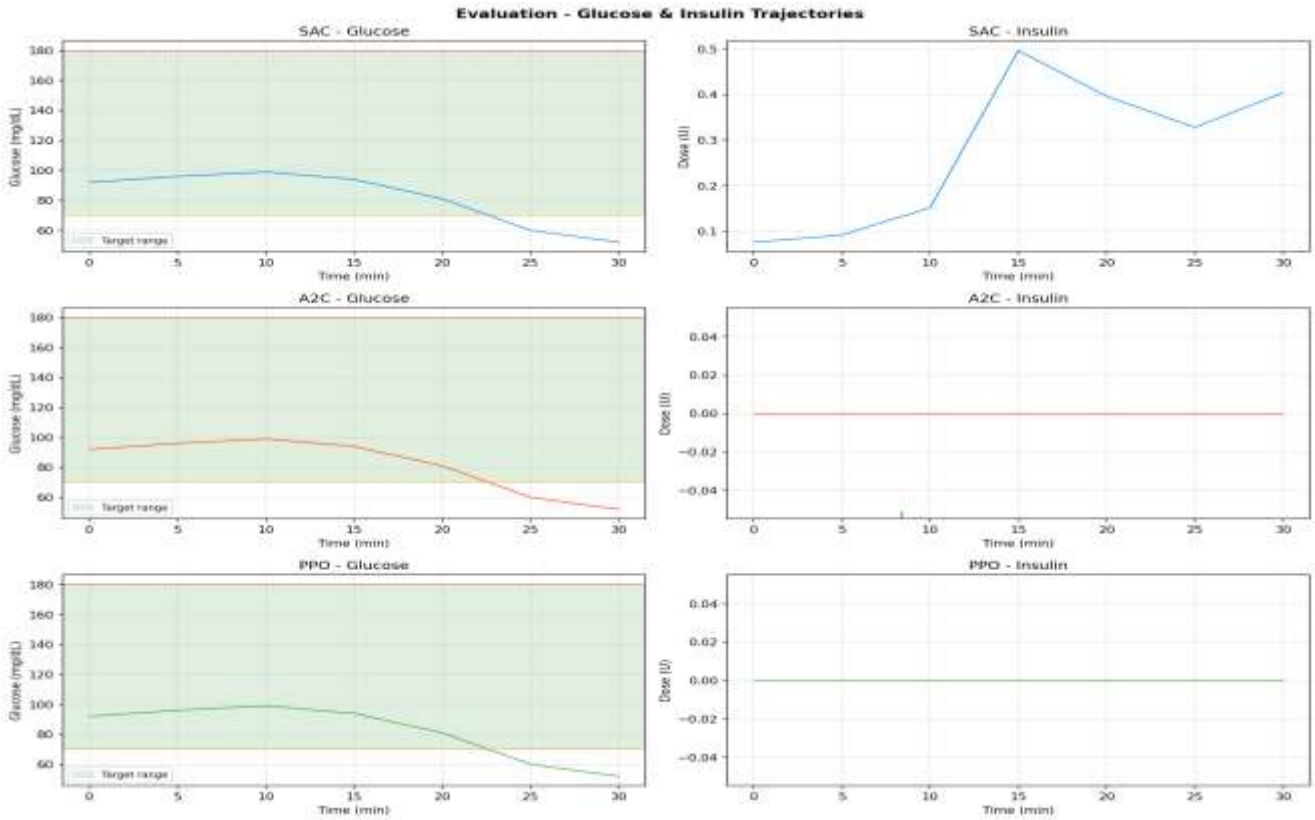

Figure 5: Representative glucose and insulin dose trajectories during policy evaluation for SAC, A2C, and PPO

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Clinical Comparison

Table 5 presents the clinical performance of SAC, A2C, and PPO based on Time in Range (TIR), hypoglycaemia rate, hyperglycaemia rate, mean glucose deviation, and average insulin dose. All three algorithms achieved identical glucose-related metrics, including a TIR of 72.7%, because these measures were computed from the same historical glucose trajectory replayed by the open-loop environment.

Table 5: Clinical Performance comparison of SAC, A2C, and PPO

Clinical Metric	SAC	A2C	PPO
Time in Range (%)	72.7	72.7	72.7
Hypoglycaemia rate (%)	1.9	1.9	1.9
Hyperglycaemia rate (%)	25.4	25.4	25.4
Mean glucose deviation (mg/dL)	45.6	45.6	45.6
Average insulin dose (U)	1.769	0.0	0.0

Although the glucose outcome metrics were identical, average insulin dose differed markedly across the algorithms. SAC maintained an average insulin dose of 1.769 U per step, whereas both A2C and PPO administered virtually no insulin during evaluation. This contrast indicates that similar clinical outcome metrics did not necessarily reflect similar treatment policies.

Figure 6 further illustrates these findings. While SAC, A2C, and PPO exhibited identical values for Time in Range, hypoglycaemia rate, hyperglycaemia rate, and mean glucose deviation, only the insulin dosing metric distinguished the learned policies. The clinical metrics, therefore should be interpreted alongside policy behaviour.

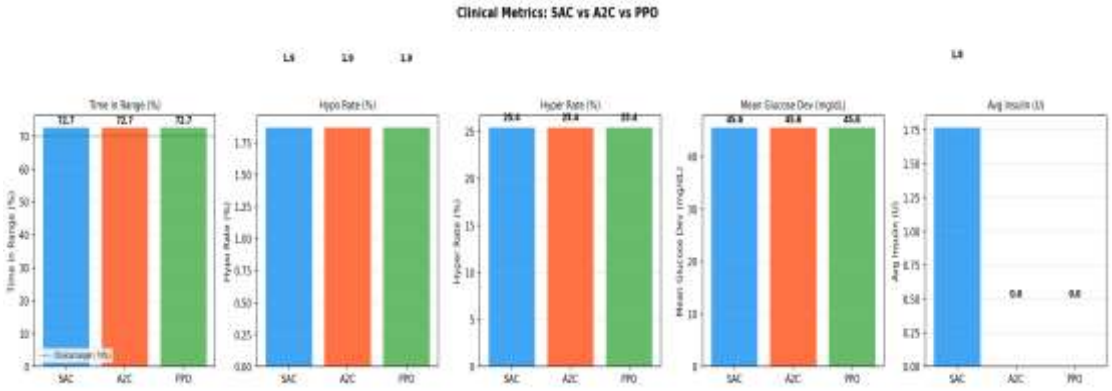

Figure 6. Clinical performance comparison of SAC, A2C, and PPO showing Time in Range, hypoglycaemia rate, hyperglycaemia rate, mean glucose deviation, and average insulin dose.

Comparative Interpretation

Collectively, the results demonstrate that cumulative reward alone is insufficient for evaluating reinforcement learning policies in open-loop clinical environments. Although A2C and PPO achieved higher rewards, both converged to near-zero insulin dosing policies, whereas SAC maintained adaptive dosing behaviour despite receiving a lower cumulative reward. These findings highlight the importance of combining reward-based, clinical, and behavioural metrics when evaluating reinforcement learning systems for personalized insulin dosing.

CONCLUSION

This study compared three reinforcement learning algorithms A2C, PPO, and SAC for personalized insulin dosing using real-world data from the OhioT1DM dataset. Although A2C and PPO achieved higher cumulative

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

rewards, both converged to near-zero insulin dosing policies that exploited the reward formulation rather than demonstrating clinically meaningful glucose management. In contrast, SAC maintained adaptive insulin dosing behaviour despite achieving a lower cumulative reward.

These findings demonstrate that cumulative reward alone is insufficient for evaluating reinforcement learning policies in open-loop clinical environments. Because historical glucose trajectories are unaffected by the agent's actions, conventional glucose outcome metrics may fail to distinguish clinically meaningful policies from those that simply exploit weaknesses in the reward design. Incorporating behavioural measures, such as insulin dosing patterns, alongside clinical outcome metrics provides a more reliable assessment of policy quality.

These results should be interpreted in light of two limitations of the present study. First, because the training and evaluation environment is open-loop by construction, replaying a single patient's historical glucose trace rather than simulating the causal effect of insulin on glucose, it cannot capture true closed-loop physiological feedback, and the reward –hacking behaviour reported here reflects this open-loop formulation rather than a general claim about closed-loop algorithm performance. Second, all three algorithms were trained using default stable-baselines3 hyperparameters without algorithm-specific tuning; while this was a deliberate choice to isolate differences attributable to learning paradigm rather than tuning effort. It leaves open whether the zero-dosing convergence observed in A2C and PPO would persist under tuned configurations. These findings demonstrate that cumulative reward alone is insufficient for evaluating reinforcement learning policies in open-loop clinical environments.

The study highlights the importance of carefully designing reward functions and evaluation strategies for healthcare reinforcement learning. More broadly, it provides evidence that behaviour-focused evaluation can improve the safety and reliability of AI-driven clinical decision-support systems by identifying undesirable policies before deployment.

Future work will address these limitations directly. A brief hyperparameter-tuning check on A2C and PPO, testing a small set of non-default configurations, will establish whether the near-zero dosing policies reported here are robust to hyperparameter choice or partly attributable to using default settings. In parallel, the current single-patient analysis will be extended across the full OhioT1DM cohort to test whether the reward-hacking pattern generalizes across individual glucose profiles, an extension that requires no changes to the underlying environment or reward design. The proposed framework will also be extended to closed-loop or hybrid simulation environments such as the UVA/Padova T1DM simulator, in which insulin actions directly influence subsequent glucose dynamics. Evaluating multiple patients, exploring reward functions informed by clinical expertise, and validating the framework on larger cohorts will further support the development of robust and clinically applicable reinforcement learning systems for personalized diabetes management.

Author Contributions

Chinatu M. Anyanwu conceived the study, designed the methodology, developed the reinforcement learning framework, implemented the experiments, analyzed the data, and drafted the manuscript. Nkiru C. Ogbonna supervised the research, provided methodological guidance, and critically reviewed the manuscript. Mary Ofuru Kama contributed to the study design, interpretation of results, and manuscript revision. Stephen Uche Udeh contributed to the experimental design, data interpretation, and technical review of the manuscript. Ogechi Gift Onyedi provided clinical insights into diabetes management, contributed to the interpretation of the findings, and reviewed the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors acknowledge the developers of the OhioT1DM dataset for making the clinical data publicly available for research. The authors also appreciate the constructive comments and suggestions from anonymous reviewers, which helped improve the quality of this manuscript.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

REFERENCES

1. Bolland, A., Lambrechts, G., & Ernst, D. (2024). Off-policy maximum entropy rl with future state and action visitation measures. arXiv preprint arXiv:2412.06655.
2. Dénes-Fazakas, L., Szilágyi, L., Kovács, L., De Gaetano, A., & Eigner, G. (2024). Reinforcement learning: a paradigm shift in personalized blood glucose management for diabetes. Biomedicines, 12(9), 2143.
3. Elsayed, N. A., Aleppo, G., Bannuru, R. R., Bruemmer, D., Collins, B. S., Ekhlaspour, L., & American Diabetes Association Professional Practice Committee. (2024). 16. Diabetes Care in the Hospital: Standards of Care in Diabetes—2024. Diabetes Care, 47.
4. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. Proceedings of the 35th International Conference on Machine Learning. https://doi.org/10.48550/arXiv.1801.01290
5. Lei, J., Sun, X., Li, Y., Li, K., Zhang, S., Zeng, H., & Zhang, Y. (2024, August). An Improved Adaptive Glucose Control Approach for Type 1 Diabetes with Temporal Dependence. In 2024 IEEE 9th International Conference on Computational Intelligence and Applications (ICCIA) (pp. 209-214). IEEE.
6. Manas, S., Pillai, G. N. & Gupta, M. K. (2023). Improved Soft Actor-Critic: Reducing Bias and Estimation Error for Fast Learning. IEEE International Student's Conference on Electrical, Electronics and Computer Science (SCEECS), 1 - 9,2023,doi:10.1109/SCEECS57921.
7. Milton T. & Lieck R. (2024). Fully-Automated Patient-Agnostic Diabetes Management with Deep Reinforcement Learning. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1085-1091.
8. Mnih, V. , Adria, P. B. , M. Mehdi, G. Alex, H. Tim, P. L. Timothy, S. David & K. Koray, (2016). Asynchronous Methods for Deep Reinforcement Learning. Proceedings of the 33rd International Conference on Machine Learning, New York. NY USA. JLMR. W & CP, 48, doi:10.48550/arXiv.1602.01783.
9. Parveen, A. (2021). A Personalized Deep Learning Approach for Blood Glucose Prediction in People with T1DM (Master's thesis, Stevens Institute of Technology).
10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347.
11. Singh, R., & Raj R. R. (2023). Optimizing Glycemic Control in Type 1 Diabetic Patients using a Deep Learning-Based Artificial Pancreas with a Secure Glucagon and Insulin Delivery System. bioRxiv, 12. doi: https://doi.org/10.1101/2023.12.07.566476.
12. Tuomas, H., Aurick, Z. Pieter, A. & Sergey, L. (2018). Soft Actor-Critic: Off - policy Maximum Entropy Deep reinforcement Learning with a Stochastic Actor. International Journal of Research and Innovation in Social Sciences, doi:10.48550/arXiv.1801.01290, https://www.researchgate.net/publication/322306636_Soft_Actor-Critic_Off-policy_Maximum_Entropy_Deep_Reinforcement_Learning_with_a_Stochastic_Actor
13. Zhao, X., Ding, S., An, Y., & Jia, W. (2019). Applications of asynchronous deep reinforcement learning based on dynamic updating weights: X. Zhao et al. Applied Intelligence, 49(2), 581-591.
14. Zheng, M., Zhang, J., Zhan, C., Ren, X., & Lü, S. (2025). Proximal policy optimization with reward-based prioritization. Expert Systems with Applications, 283, 127659.