Offline Reinforcement Learning for Safer Blood Glucose Control in People with Type 1 Diabetes

Journal of Biomedical Informatics
Citations: 58 | Published: June 2023

Saman Khamesian

05/14/2025

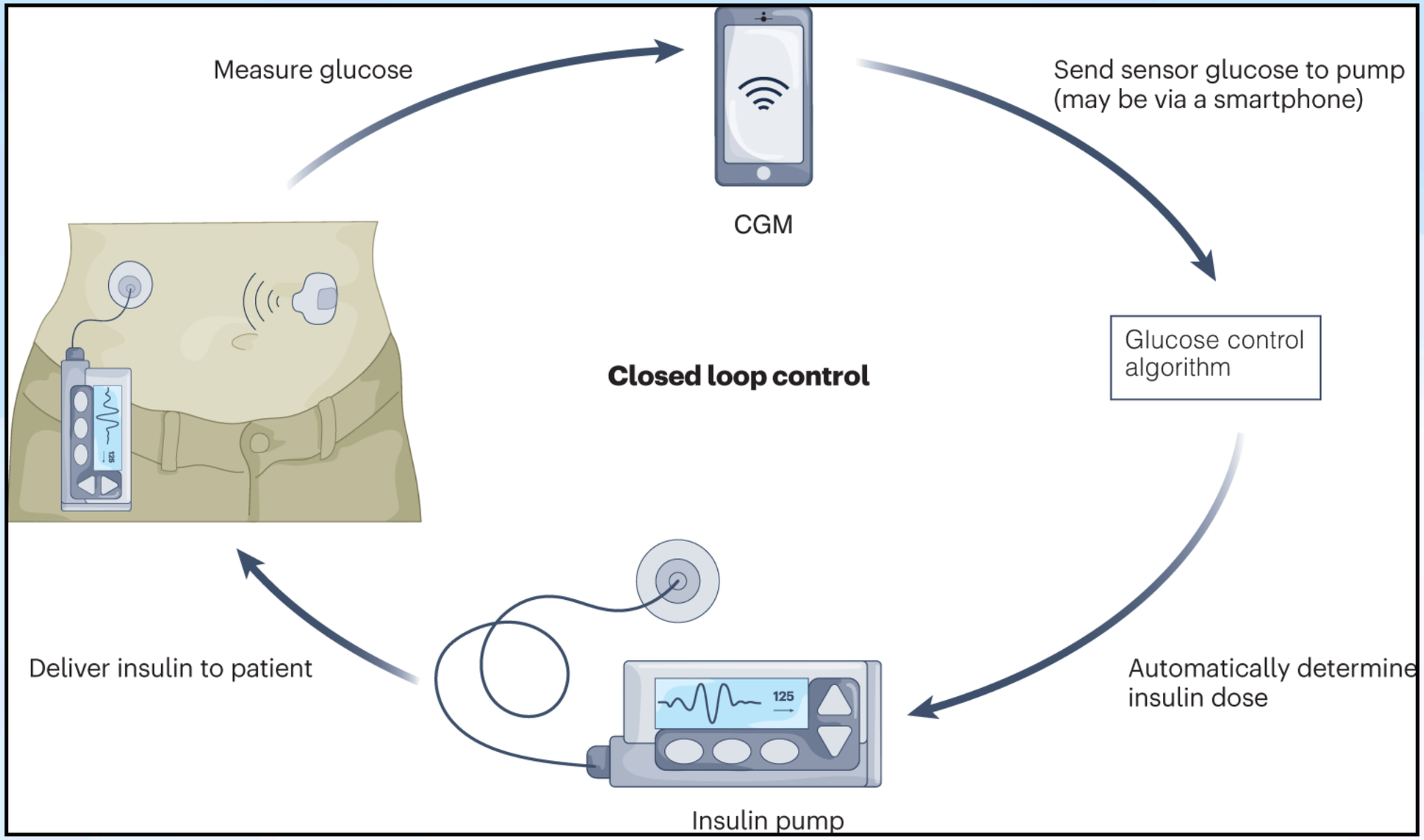
INTRODUCTION

- People with T1D must regularly monitor their blood glucose levels to:
  - Estimate the correct dosage of insulin
  - Estimate carbohydrate intake
  - Avoid hyperglycemia and hypoglycemia

- Hybrid closed loop systems provide an opportunity for people with T1D to automatically regulate their basal insulin dosing.

- These devices consist of:
  - **CGM Device:** Measures the blood glucose level of the user
  - **Insulin Pump:** Including a control algorithm that uses the CGM data to instruct the insulin pump to deliver the required dosage.

INTRODUCTION



INTRODUCTION

- Most HCL devices use these algorithms to control insulin delivery:
  - PID (Predictive integral derivative) controllers
  - MPC (Model Predictive Control) algorithms
- These methods are:
  - ✅ Robust
  - ✅ Easily interpretable
  - ❌ Limited in performance and adaptability

INTRODUCTION

- PID limitations:
  - Often overestimate insulin doses after meals
  - Cannot easily integrate factors like:
    - Insulin activity
    - Time of day
    - Physical exercise
- MPC limitations:
  - Rely on linear or simplified models
  - Fail to capture the full complexity of glucose dynamics

INTRODUCTION

- **How can we build a practical controller?**

- **Reinforcement Learning (RL)** teaches an agent to choose the best sequence of actions to maximize a reward — making it ideal for complex, goal-driven tasks like managing blood glucose.

INTRODUCTION

- **Most current approaches** use **online RL**, which means:
  - Agent needs to interact with either a real patient or a simulator during training
  - It improves through a trial-and-error process — It's risky and slow
- **This study explores offline RL** is a safer alternative, because:
  - The agent does not explore or interact with the environment during training.
  - It learns entirely from a **pre-collected dataset** of previous decisions and outcomes, avoiding the risks of unsafe experimentation.

METHODOLOGY

- **How can we define our problem in RL context?**

- Glucose control in HCL systems can be understood as a **step-by-step decision-making process**, where each decision affects the future.

- It's modeled as a **Partially Observable Markov Decision Process.**

- At every moment (or time-step) t:
  - The agent is in a certain **state** $s \in S$ (e.g., based on blood glucose level).
  - It takes an **action** $a \in A$ (like giving insulin).
  - It gets a **reward** $r = R(s)$ based on how good that action was.
  - Then it moves to a **new state** $s' \in S$ depending on the action it just took.

METHODOLOGY

$$s_t = [g_t, g_{t-1}, g_{t-2}, \cdots, g_{t-8}, I_t, M_t]$$

- g_t is the current glucose value
- g_{t-8} is the glucose value 4 hours earlier (Values are sampled every 30 minutes over a 4-hour window)
- I_t **Insulin-on-board:** Combining recent basal and bolus insulin
- M_t **Carbohydrate activity:** Reflecting recent carb intake
- This representation simplifies true insulin and carbohydrate activity by assuming that they decay linearly to zero over a four hour period.

METHODOLOGY

- The reward for the agent was given by the negative of the Magni risk function, which models the clinical risk for a given blood glucose value :

$$risk(g_t) = 10 \cdot \left(3.5506 \cdot (\log(g_t)^{0.8353} - 3.7932)\right)^2$$

- An additional penalty of **-10,000** was added when blood glucose goes outside of **10 to 1,000 mg/dL** to discourage the agent from ending the episode on purpose just to avoid future negative rewards.

- The reward is at a maximum when blood glucose is approximately in the centre of the target range (70-180 mg/dl).

METHODOLOGY

- In this work they used three different model-free offline RL approaches and compared them with the baselines.
  - **Batch Constrained Deep Q-learning (BCQ)**
  - **Conservative Q-learning (CQL)**
  - **Twin Delayed DDPG with Behavioral Cloning (TD3-BC)**

METHODOLOGY

- The performance of the offline RL methods were compared to two baseline algorithms:
  - **A tuned PID controller:** The PID algorithm is a robust classical control mechanism capable of correcting for both short and long-term deviations from a target blood glucose value.
  - **An online RL method, recurrent soft actor-critic (SAC-RNN):** The method used in this paper is a variation of this method, using long-short-term memory (LSTM) layers in place of gated recurrent unit (GRU) layers.

DATASET

- The study used the **UVA/Padova T1D simulator** to generate the data.
- This simulator models the **human metabolic system**, including glucose-insulin dynamics, and is designed to replace pre-clinical trials.
- It provides full control over the **size and quality** of the dataset, which is useful for training RL agents.
- The simulator includes a virtual cohort of 30 patients: 10 children, 10 adolescents, and 10 adults.
- Each virtual patient has **individualized responses** to:
  - Meal carbohydrates
  - Basal and bolus insulin
  - Interactions with CGM and insulin pump devices

DATASET

- They simulated 208 days of glucose activity to collect 100,000 samples for training the RL models.
- Individually each patient's data was used to train each algorithm across three seeds.
- **Noise was added** to the simulated data to encourage more realistic exploration.

RESULTS

Metric	Description	Target
Time-in-Range (TIR)	The percentage time for which blood glucose measurements fall within the healthy glucose range (70-180 mg/dl). Increased TIR is strongly associated with a reduced risk of developing micro-vascular complications [56].	>70%* [57]
Time-Below-Range (TBR)	The percentage time for which blood glucose measurements fall in the low blood glucose range (<70 mg/dl). Combined with TIR, this can additionally act as an indirect measure of time spent in the high blood glucose range (>180 mg/dl).	<4% [57]
Coefficient of Variation (CV)	The relative dispersion of blood glucose values around their mean. Increased CV is linked to an elevated risk of severe low blood glucose events (<54 mg/dl) and vascular tissue damage [58].	<36% [59]
Failure	The percentage of test rollouts in which blood glucose levels reached values <10 mg/dl or >1000 mg/dl. For context, blood glucose <40 mg/dl is considered life-threatening and can result in major cardiovascular and cerebrovascular problems [60].	0%

Table 1: A description of the glycemic control metrics utilised for evaluation alongside their clinically recommended values. *For ages <25 this target decreases to >60%.

RESULTS

Algorithm	Reward	TIR (%)	TBR (%)	CV (%)	Failure (%)
BCQ†	-41,034 ± 1,060	**65.8 ± 0.6**	1.0 ± 0.1	35.1 ± 0.4	**0.00**
CQL†	-45,259 ± 1,071	56.2 ± 0.5	**0.1 ± 0.1**	30.3 ± 0.3	**0.00**
TD3-BC†	**-37,955 ± 547**	65.3 ± 0.5	0.2 ± 0.1	33.3 ± 0.2	**0.00**
SAC-RNN‡	-93,480 ± 71,826	34.9 ± 3.1	4.1 ± 0.7	**29.6 ± 1.3**	13.3
PID§	-49,077 ± 556	61.6 ± 0.3	0.4 ± 0.1	33.5 ± 0.2	**0.00**

- The offline RL algorithm TD3-BC achieved the best performance.
- In addition, the algorithm yields a **3.7±0.6% increase to TIR** and a reduction in TBR from 0.4% to 0.2%.
- This equates to almost **an additional hour per day** in which patients would experience an improved quality of life.

RESULTS

(a) Adults (aged 26 to 68 years).

Algorithm	Reward	TIR (%)	TBR (%)	CV (%)	Failure (%)
BCQ†	**-17,445 ± 290**	**72.6 ± 0.6**	0.1 ± 0.0	25.6 ± 0.2	**0.00**
CQL†	-20,748 ± 301	62.5 ± 0.5	**0.0 ± 0.0**	**22.9 ± 0.2**	**0.00**
TD3-BC†	-19,538 ± 381	70.0 ± 0.6	0.1 ± 0.1	26.0 ± 0.2	**0.00**
SAC-RNN‡	-49,600 ± 6,075	42.5 ± 3.4	5.1 ± 0.1	26.0 ± 1.5	6.6
PID§	-19,783 ± 262	65.8 ± 0.3	**0.0 ± 0.0**	24.4 ± 0.2	**0.00**

(b) Adolescents (aged 14 to 19 years).

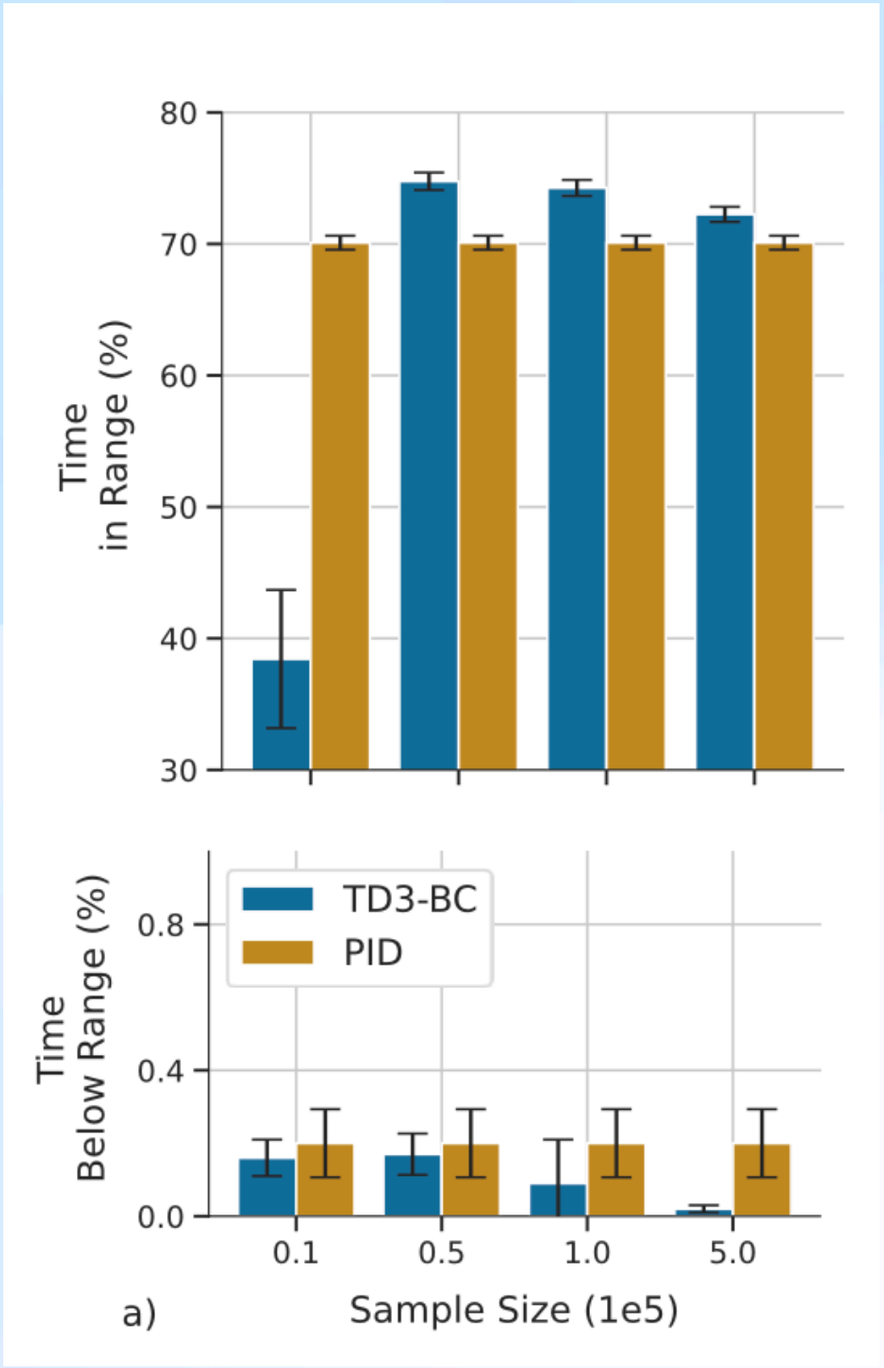
Algorithm	Reward	TIR (%)	TBR (%)	CV (%)	Failure (%)
BCQ†	-39,932 ± 347	**64.9 ± 0.3**	1.6 ± 0.1	33.28 ± 0.3	**0.00**
CQL†	-43,847 ± 389	56.12 ± 0.4	**0.1 ± 0.0**	27.6 ± 0.3	**0.00**
TD3-BC †	**-39,363 ± 614**	62.0 ± 0.6	**0.1 ± 0.1**	**26.1 ± 0.3**	**0.00**
SAC-RNN‡	-80,102 ± 5,597	25.9 ± 3.2	2.7 ± 0.9	**26.1 ± 0.2**	13.3
PID§	-40,180 ± 465	60.6 ± 0.2	0.1 ± 0.0	30.75 ± 0.2	**0.00**

(c) Children (aged 7 to 12 years).

Algorithm	Reward	TIR (%)	TBR (%)	CV (%)	Failure (%)
BCQ†	-61,374 ± 2,543	56.9 ± 0.9	1.2 ± 0.2	41.9 ± 0.6	**0.00**
CQL†	-66,346 ± 2,522	44.2 ± 0.7	**0.3 ± 0.1**	37.2 ± 0.4	**0.00**
TD3-BC †	**-51,713 ± 646**	**60.1 ± 0.4**	**0.3 ± 0.1**	40.4 ± 0.2	**0.00**
SAC-RNN‡	-97,760 ± 6,339	37.3 ± 2.8	4.0 ± 1.0	**36.0 ± 1.4**	20.0
PID	-57,700 ± 941	54.2 ± 0.4	1.7 ± 0.1	41.6 ± 0.3	**0.00**

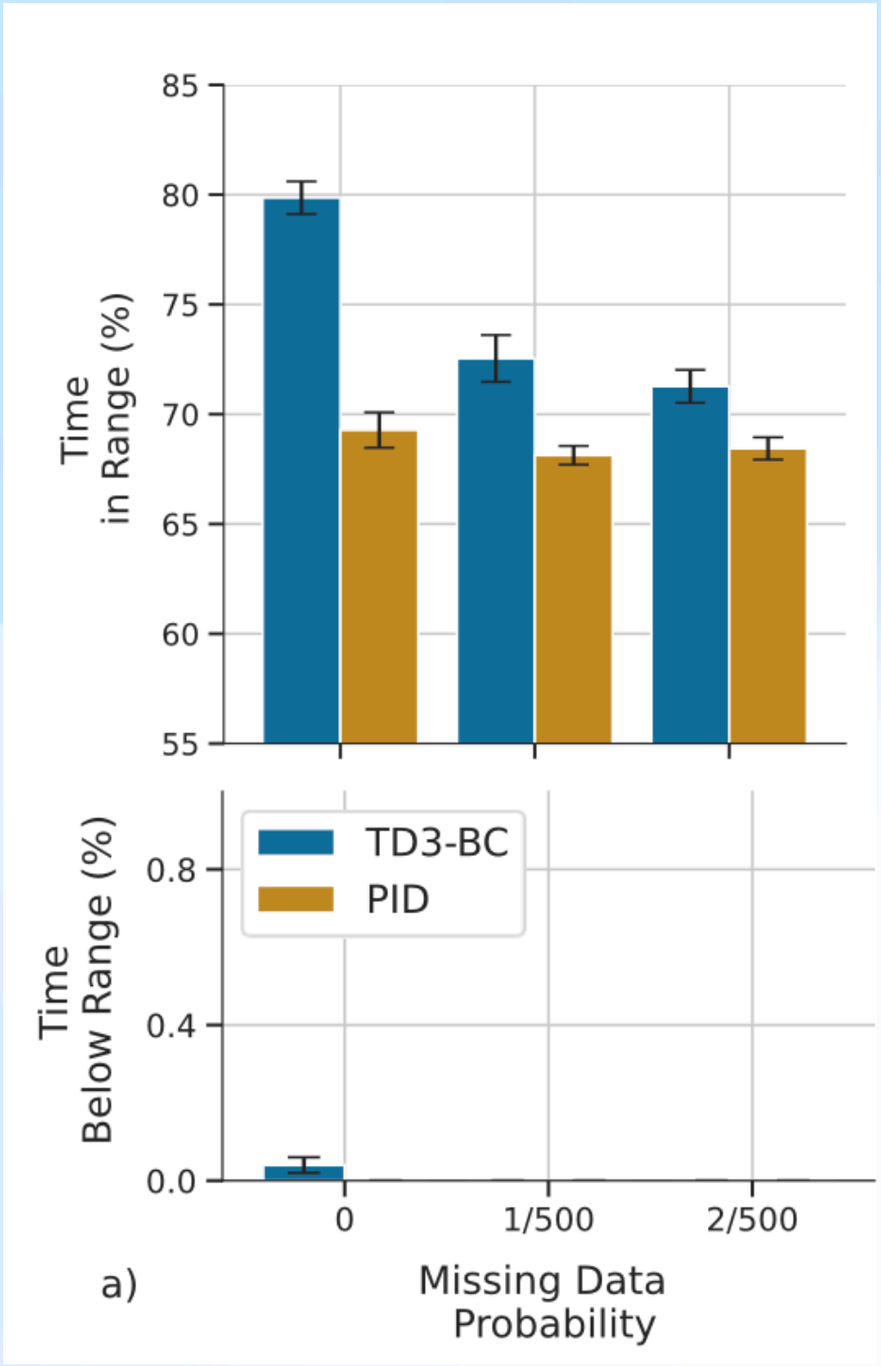
- A breakdown of glucose control performance when divided by patient age

RESULTS



- Shows the effect of varying sample size on the performance of the offline RL algorithm TD3-BC vs. PID.
- For the greatest number of samples $5e^5$ (>1,000 days), TD3-BC achieves a 2.2±1.1% increase to TIR and a 0.2 ±0.1% reduction to TBR.
- This result is significant for the application of offline RL to future hybrid closed loop systems as 100 days represents a feasible timescale for data collection in patient populations.

RESULTS



- Shows the effect of missing data on the performance of TD3-BC.
- The reduction in performance associated with the addition of missing data is particularly significant, resulting in a decrease of 7.32 ±1.3% to TIR.

LIMITATIONS

- The most significant limitation of the presented evaluation is the use of the T1D simulator.
- These environments only capture a fraction of the complexity involved in realistic blood glucose dynamics; neglecting events such as stress, activity and illness.
- The effect of actions may only become apparent in some instances over several days, requiring blood glucose to be modeled over significantly longer prediction horizons than are currently deployed in commercial systems.

Thank you for your attention