National College of Ireland

Hybrid Reinforcement Learning for Personalized Diabetes Care

MSc Research Project
MSc in Artificial Intelligence

Shayshank Rathore
Student ID: 23348186

School of Computing
National College of Ireland

Supervisor: Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shayshank Rathore
Student ID:	23348186
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Abdul Shahid
Submission Due Date:	01/09/2025
Project Title:	Hybrid Reinforcement Learning for Personalized Diabetes Care
Word Count:	6578
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Shayshank Rathore
Date:	31st August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid Reinforcement Learning for Personalized Diabetes Care

Shayshank Rathore
23348186

Abstract

Custom diabetes care needs decision support that adapts quickly to each patient while remaining transparent to clinicians. We face major challenge with a lack of explainability in action of Artificial Intelligence (AI) and a heavy dependence on large amounts of data for training in medical field. We present a hybrid reinforcement learning (RL) approach that combines reward decomposition. It separating clinically meaningful objectives such as glycaemic control, avoidance of adverse events and intervention burden with a lightweight meta-learning routine for fast per-patient adaptation. Using a custom Gymnasium environment derived from the UCI diabetes time-series, we train Proximal Policy Optimization (PPO) and compare it to a rule-based controller. Evaluation focuses on Time-in-Range (TIR), hypoglycaemic events, insulin usage and sample-efficiency. In our current configuration, PPO eliminates hypoglycaemic steps and reduces insulin use by $\sim$86% relative to the rule baseline. A pilot meta-learning step yields small per-task reward gains, indicating potential for personalization with limited data. We discuss why control remains suboptimal (coarse action space, reward balance, largely deterministic dynamics) and outline concrete remedies. Overall, the method delivers interpretable trade-offs and data-efficient adaptation, offering a pragmatic path toward trustworthy RL-based decision support in diabetes care.

Keywords: Reinforcement Learning, Personalized Diabetes Care, Reward Decomposition, Meta-Learning, Proximal Policy Optimization (PPO)

1 Introduction

1.1 Background

Diabetes mellitus is a chronic metabolic disorder characterised by impaired insulin production or action, leading to persistently elevated blood glucose and substantial morbidity worldwide Shaikh et al. (2022). Effective management is inherently individualised treatment plans should reflect a person's age, comorbidities, lifestyle and response to prior therapy. Yet in routine practice, care often follows uniform protocols limiting responsiveness to patient heterogeneity.

Artificial Intelligence (AI) and in particular Reinforcement Learning (RL) offers a way to tailor decisions over time by learning policies that optimise long-term clinical outcomes from patient trajectories Ali (2022). However, two barriers have limited clinical uptake. First, many RL methods are opaque, making it difficult for clinicians to understand or trust why a medical recommendation was made. Second, RL models are typically data-hungry, which is problematic in healthcare settings where labelled longitudinal data are scarce and costly to obtain Yu et al. (2021).

This study addresses both issues by combining reward decomposition which, separates overall reward into clinically meaningful components such as glycaemic control, adverse-event avoidance and treatment burden with meta-learning to improve data efficiency and enable rapid adaptation to new patients. We construct a custom Gymnasium environment derived from the UCI diabetes dataset to simulate sequential treatment decisions and evaluate transparent adaptive RL policies. The goal is to demonstrate a practical pathway to explainable and sample-efficient decision support for personalized diabetes care.

1.2 Problem Statement and Research Gaps

More than 853 million adults live with diabetes, creating a heavy clinical and economic burden Federation (2025). Care must be individualized because disease trajectories and treatment responses vary widely. This heterogeneity makes day-to-day decisions hard to standardize. Reinforcement Learning (RL) can help by learning treatment policies that adapt over time to each patient Banumathi et al. (2025).

Two barriers limit clinical use. First, many RL models are not explainable. Clinicians cannot see why an action was chosen, which reduces trust Abdellatif et al. (2023). Second, RL is often data-hungry because the learning process require a large number of interactions with the environment to explore and exploit different states and actions effectively. Healthcare datasets are small and noisy also patient-specific data are scarce and costly to obtain. These factors slow learning and reduce reliability.

Recent work on reward decomposition improves transparency by splitting the objective into clinical terms such as efficacy, side effects and cost Hu et al. (2023). Meta-learning can improve data efficiency by enabling rapid adaptation from limited patient data. However, most studies treat these aims separately. Few combine reward decomposition and meta-learning for diabetes management.

This study addresses that gap. We implement an RL framework that pairs reward decomposition with meta-learning to provide more explainable and more adaptive treatment recommendations for diabetes management.

1.3 Research Aim and Objectives

Research Aim

Design and evaluate a practical reinforcement learning framework that combines reward decomposition for transparent decisions and meta-learning for fast per-patient adaptation in personalized diabetes care with a focus on clinical safety and sample efficiency.

Research Objectives

- Build a custom Gymnasium environment from the UCI diabetes data for sequential treatment decisions.
- Implement a PPO-based agent with reward decomposition to expose trade-offs between glycaemic control, adverse events and intervention burden.
- Added a lightweight meta-learning routine to enable rapid adaptation to new patient profiles.
- Evaluate against a rule-based baseline using clinical and RL metrics: Time-in-Range (TIR), hypoglycaemic events, insulin use per step, learning curves and sample efficiency.
- Assess explainability via per-channel reward logs and action distributions also report limitations and ethical considerations for clinical translation.

1.4 Overview of Methodology

We build a practical Reinforcement Learning (RL) pipeline that couples reward decomposition with meta-learning for personalised diabetes care.

Data & Preprocessing:- We use the UCI diabetes time-series based dataset. We clean missing values align records into fixed steps normalise features and split by patient into train/validation/test.

Environment:- A custom Gymnasium environment encodes:

- **State:** Recent glucose and treatment history and available covariates.
- **Action:** Three discrete insulin-intensity choices.
- **Reward:**
    1. Glycaemic improvement
    2. Hypo / Side-effect penalty
    3. Treatment cost

Agent and Baseline:- The main agent is PPO (Stable-Baselines3). A rule-based controller serves as the baseline.

Meta-learning:- A lightweight per-patient adaptation step fine-tunes the PPO policy on small task batches to improve data efficiency and personalisation.

Training:- We fix random seeds, use early stopping on validation signals and save models, logs and plots for reproducibility.

Evaluation:- We report Time-in-Range (TIR), hypoglycaemic and hyperglycaemic steps, insulin use per step, average return and sample efficiency. We assess generalisation on unseen patients.

Explainability:- We analyse reward channels and action distributions over states to show why the agent acts and how it trades control, safety and burden.

2 Literature Review

2.1 RL for Personalized Treatment

Reinforcement Learning (RL) is a natural fit for sequential clinical decisions. It learns policies that optimise long-term outcomes from patient trajectories. In diabetes care, RL has been used to recommend insulin dosing and drug adjustments, with signs of improved glycaemic control over fixed protocols Sun et al. (2021). Yet adoption remains limited in practice.

2.2 Barriers: Transparency and Trust

Clinicians need to understand why an action was chosen. Many RL models are black boxes, which weakens trust and slows clinical uptake Abdellatif et al. (2023). Lack of interpretability also hinders safety review and governance.

2.3 Barriers: Data Scarcity and Heterogeneity

RL is often data-hungry. Healthcare data are sparse, noisy and expensive to collect. Patient heterogeneity further fragments the data and reduces sample efficiency Yu et al. (2021).

2.4 Reward Decomposition for Explainability

Reward decomposition splits a composite reward into clinical parts such as efficacy, adverse events and treatment burden. Vouros (2023) formalises this approach for explainable deep RL, while Hu et al. (2023) demonstrate interpretable RL in a clinical ICU setting.

2.5 Meta-Learning for Fast Adaptation

Model-agnostic meta-learning methods such as MAML Finn et al. (2017) and PEARL Rakelly et al. (2019) can adapt policies to new patients using few trajectories. Rafiei et al. (2024) report rapid adaptation on healthcare problems and review meta-RL for personalised medicine.Banumathi et al. (2025).

2.6 Missing Piece: Combining Both Aims

Most studies emphasise either interpretability or fast adaptation. Few unite reward decomposition with meta-learning in diabetes care. This limits real-world impact, where clinicians need transparent reasoning and rapid personalisation.

2.7 Related Approaches Beyond RL

Supervised recommenders (e.g., gradient boosting, random forests) can map features to clinician-approved actions and are compatible with SHAP/LIME for feature-level explanations. They do not model state transitions, so they lack adaptive planning over time Abdellatif et al. (2023). Physiological simulators (e.g., Bergman minimal model, UVA/Padova) support in-silico testing and safety checks, but rely on fixed parameters that may not generalise to broader populations. Model Predictive Control (MPC) optimises actions under constraints and offers clear safety guarantees, yet it requires accurate patient-specific models and can struggle with noisy and sparse data.
Bayesian personalised medicine encodes prior knowledge and quantifies uncertainty, but scaling to high-dimensional, long-horizon decisions is challenging Hu et al. (2023).
Hybrid ideas exist adding safety constraints to RL or using Bayesian priors but a principled, end-to-end design that is both transparent and rapidly adaptive for diabetes remains under-explored.

2.8 Positioning of this Study

We address the gap by pairing reward decomposition (for clinical transparency) with meta-learning (for fast, data-efficient adaptation). The goal is a practical RL framework that clinicians can inspect, audit and adapt to diverse patient profiles.

3 Dataset and Preprocessing

3.1 Dataset

We use the UCI Diabetes time-series dataset originally prepared for the AIM '94 community Kahn (1994). It contains records for 70 patients with insulin-dependent diabetes. Each record consists of four tab-separated fields: `Date` (MM-DD-YYYY), `Time` (HH:MM), an event `Code` and a numeric `Value`. The `Code` specifies the type of event. For example, insulin doses (33 = Regular, 34 = Neutral Protamine Hagedorn (NPH), 35 = UltraLente) and blood-glucose measurements (e.g., 58 = pre-breakfast, 60 = pre-lunch). The `Value` field represents either the insulin units for dose events or the glucose level Milligrams per Deciliter(mg/dL) for measurement events.

Data came from two sources: some patients used an electronic recorder with a real clock, while others logged events on paper using "logical" times (breakfast, lunch, dinner, bedtime). For paper logs, the repository maps these to fixed times (08:00, 12:00, 18:00, 22:00). This mix produces irregular sampling and missingness patterns typical of real-world logs. Although the dataset is relatively good for exploration, it is rich in event types, making it suitable for a Reinforcement Learning (RL) environment focused on action transparency and fast adaptation. It is not a large clinical registry and should

Study	Setting / Method	Key Contribution	Limitation	Relevance
Sun et al. (2021)	RL for Type-2 diabetes treatment	RL can outperform fixed protocols in glycaemic control	Limited interpretability	Motivates RL in diabetes
Abdellatif et al. (2023)	Review of RL in healthcare	Highlights transparency, safety, trust issues	No combined solution	Justifies explainability need
Vouros (2023)	Explainable Deep RL	Reward decomposition for interpretability	No fast adaptation	Provides transparency tool
Hu et al. (2023)	Interpretable RL for ICU	Aligns policies with clinical reasoning	ICU focus; no meta-learning	Shows interpretable design patterns
Rafiei et al. (2024)	Meta-learning in healthcare	Rapid adaptation from limited data	No reward decomposition	Provides adaptation tool
Banumathi et al. (2025)	Meta-RL for personalised medicine (review)	Surveys personalisation via meta-RL	Lacks integrated interpretability	Supports hybrid rationale

Table 1: Summary of related studies contributions and limitations.

not be treated as such. Instead, we use it to simulate sequential decision-making and to study explainability and data-efficiency under realistic constraints.

3.2 Preprocessing

The goal was to transform irregular event logs into clean, patient-wise sequences suitable for reinforcement learning (RL) agent training. The preprocessing pipeline consisted of several stages:

Ingestion and Parsing: The 70 patient files were loaded with a persistent `patient_id`. `Date` and `Time` fields were combined into a single timestamp. Strict parsing with error coercion was applied and rows with invalid or missing timestamps were dropped. Records were then sorted by `patient_id` and timestamp.

Event Harmonisation: The UCI `Code` field was mapped to human-readable categories (e.g., glucose reading, insulin type). Only events relevant to glycaemic control and insulin delivery were retained. Duplicate rows and timestamp collisions were removed.

Temporal alignment: Logs were resampled within each patient to a fixed step size (e.g., hourly). When multiple events occurred within the same bin, aggregation rules were applied: the latest glucose reading was retained, and insulin doses were summed. Short gaps were forward-filled within a small window; longer gaps remained missing and were subsequently excluded.

Type safety and ranges: Numeric fields were coerced using `pandas.to_numeric`. Non-numeric values were discarded. Robust outlier handling was performed via quantile clipping, limiting extreme values without suppressing real variability.

Feature construction: Derived features were created to summarise recent patient history including:

- current glucose and lag values
- short rolling statistics of glucose
- insulin given in recent time windows
- time-of-day indicators

These features formed the state vector s_t. A shifted copy produced s_{t+1}, enabling construction of transition tuples.

Targets for rewards: Helper flags were created for environment use, including:

- hypoglycaemia and hyperglycaemia indicators
- insulin burden per step

Reward computation itself was deferred to the environment, which referenced these derived columns.

Normalisation:- A `StandardScaler` was fitted only on the training split and applied to validation and test sets to prevent data leakage. The fitted scaler was preserved for consistent transformation at evaluation.

Splitting:- Patients were partitioned into train, validation and test sets, ensuring no patient appeared in more than one split. This tested model generalisation to unseen patients.

Quality checks:- Final checks confirmed no missing values in key features, validated distributions and time-series trends and verified sufficient sequence lengths for rollouts.

4 Methodology

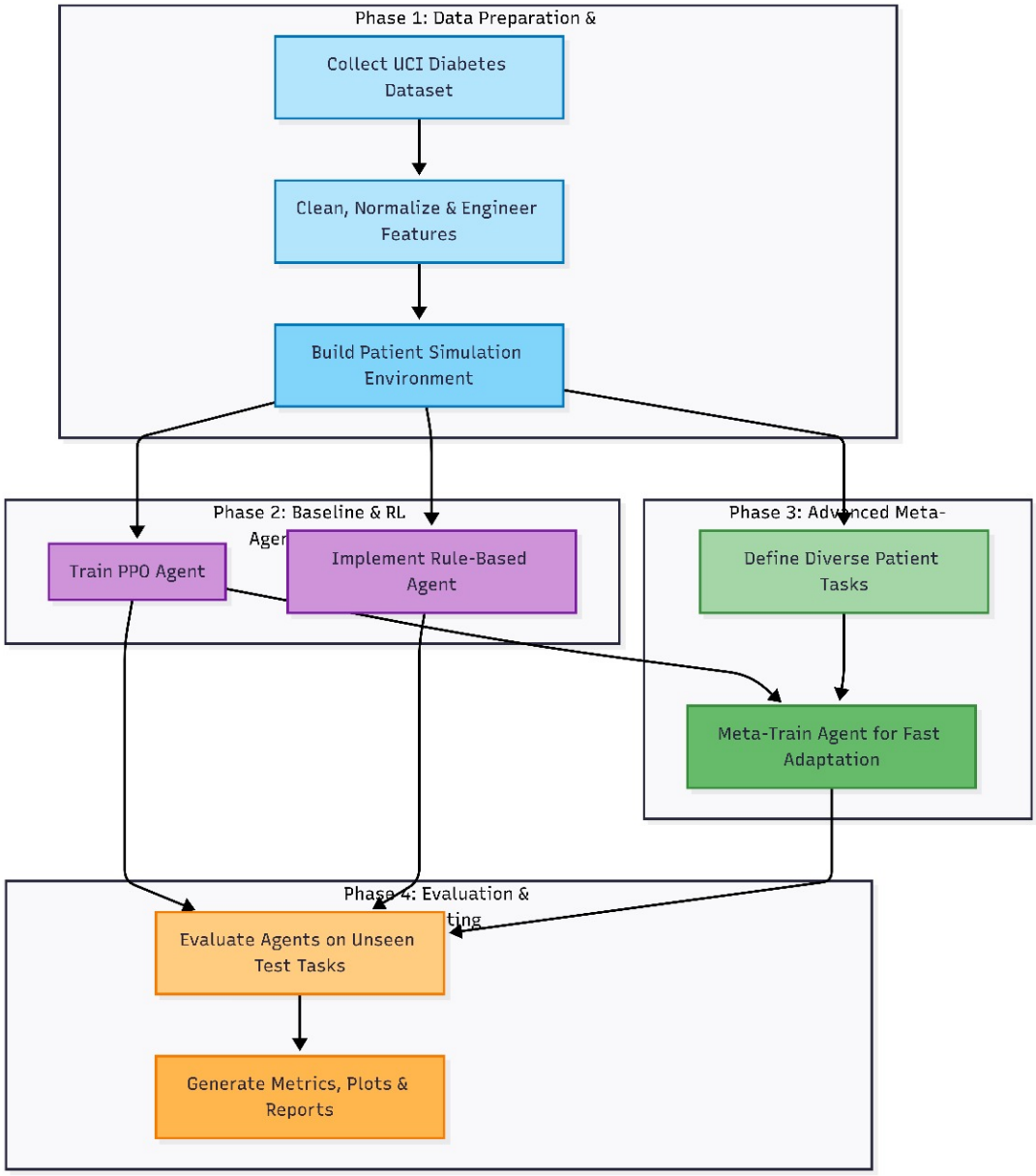

Figure 1: Stages of Methodology

4.1 Data Preprocessing and Feature Engineering

We converted the UCI Diabetes event logs into clean, patient-wise sequences suitable for reinforcement learning (RL).

Parsing: The 70 patient files were loaded with a stable `patient_id`. `Date` and `Time` were merged into a single timestamp. Rows with invalid or missing timestamps were dropped, and records were sorted by `patient_id` and timestamp.

Event mapping: UCI `Code` values were mapped to human-readable labels (e.g., glucose reading types, insulin types). Events unrelated to glycaemic control and insulin delivery were removed. Duplicate rows and exact timestamp collisions were excluded.

Temporal alignment: Each patient's records were resampled to a fixed step size (e.g., hourly). Within each bin, the latest glucose reading was retained and insulin doses were summed. Short gaps were forward-filled within a small window, while longer gaps were left missing and excluded from rollouts.

Feature set: The state vector included current glucose, short lags, rolling statistics, insulin in recent time windows and time-of-day indicators.

Type safety and ranges: Numeric fields were coerced using `pandas.to_numeric`, and extreme outliers were clipped via quantile thresholds to reduce instability.

Normalisation: A `StandardScaler` was fitted only on the training data, then applied consistently to validation and test splits.

Splitting: Patients were partitioned into training, validation and test sets to ensure evaluation on unseen patients.

Diagnostics: Final checks confirmed no missing values in model inputs and verified sufficient per-patient sequence lengths.

4.2 Environment Modeling

The problem was modeled as a Markov Decision Process (MDP):

- **State s_t.** Feature vector comprising recent glucose history, insulin events and time indicators.
- **Action a_t.** Three discrete insulin-intensity choices: none/low, moderate or high.
- **Reward r_t.** Decomposed into:
  - r_{range}: progress toward the target glucose band (Time-in-Range objective),
  - r_{hypo}: penalty for hypoglycaemia risk/events,
  - r_{cost}: treatment burden (insulin use).

Each reward channel was logged per step and episode for transparency. A learned glucose-transition model was built from the preprocessed sequences, governed state transitions and captured empirical variability.

4.3 Reinforcement Learning Framework

The primary agent was Proximal Policy Optimization (PPO), implemented using Stable-Baselines3.

Training:- Random seeds were fixed. Models were trained for approximately 30k timesteps per run using standard PPO hyperparameters (discount factor, clipping, entropy regularisation). Checkpoints, logs and plots were stored.

Reward decomposition:- PPO optimised the summed reward, but individual reward channels were logged separately for interpretability.

Meta-adaptation:- A lightweight per-patient fine-tuning step (Reptile-style) was applied to adapt PPO policies to small patient-specific tasks. Per-task reward deltas were reported as a pilot indicator of personalisation.

Reproducibility:- All code, scalers, environment configurations and random seeds were saved with each run.

4.4 Baselines and Evaluation

Baseline:- A rule-based proportional controller increased insulin under hyperglycaemia and tapered doses at lower glucose levels.

Metrics:-

- Clinical: Time-in-Range (TIR), hypoglycaemic steps, hyperglycaemic steps, insulin units per step.
- RL: Average return, Sample-efficiency (Reward progression vs Timesteps).

Explainability:- Episode-level summaries of reward channels and action-distribution plots were used to illustrate how the agent balanced control, safety and treatment burden.

Protocol:- Models were tuned on validation data and evaluated on held-out patients. Results were reported as means and distributions across episodes, with qualitative trajectory comparisons included.

5 Results and Analysis

5

5.1 Training Dynamics and Performance Metrics

We trained a Proximal Policy Optimization (PPO) agent in the custom Gymnasium environment for approximately 30k timesteps using standard Stable-Baselines3 settings ($\gamma = 0.99$, batch size = 64). A rule-based proportional controller served as the baseline. Models were selected on validation and then evaluated on held-out test patients.

Held-out test set (30 episodes).

Metric (mean over 30 eps)	PPO (test)	Rule (test)	PPO (val)	Rule (val)
Time-in-Range (TIR (%))	1.39	1.39	0.00	0.00
Hypoglycaemic steps (per ep)	0.00	1.00	0.00	1.00
Hyperglycaemic steps (per ep)	71.0	70.0	72.0	71.0
Insulin units / step	1.34	9.72	1.20	9.82
Mean glucose (mg/dL)	593.2	–	596.3	–

Table 2: PPO vs rule baseline on validation and test splits (unseen patients).

- Time-in-Range (TIR, %): PPO = 1.39, Rule = 1.39 (no improvement).
- Hypoglycaemic steps (mean per episode): PPO = 0.00, Rule = 1.00 → PPO avoids hypos.
- Hyperglycaemic steps (mean per episode): PPO = 71.0, Rule = 70.0 (both very high).
- Insulin units per step (mean): PPO = 1.34, Rule = 9.72 → ∼86% less insulin with PPO.
- Mean glucose (mg/dL): PPO = 593.2 (elevated).

Validation set (30 episodes).

- TIR (%): PPO = 0.00, Rule = 0.00.
- Hypoglycaemic steps (mean per episode): PPO = 0.00, Rule = 1.00.
- Hyperglycaemic steps (mean per episode): PPO = 72.0, Rule = 71.0.
- Insulin units per step (mean): PPO = 1.20, Rule = 9.82.
- Mean glucose (mg/dL): PPO = 596.3 (elevated).

Takeaway: The PPO agent learns a conservative, hypo-averse policy that drastically reduces insulin usage (∼86% reduction compared to the rule baseline) while eliminating hypoglycaemic events. However, it does not improve Time-in-Range under the current environment and action design. This outcome aligns with the action histogram, which shows high-intensity dosing used sparingly. The results suggest that the current reward balance and action granularity favour safety and parsimony over aggressive glucose correction.

5.2 Reward Decomposition Analysis

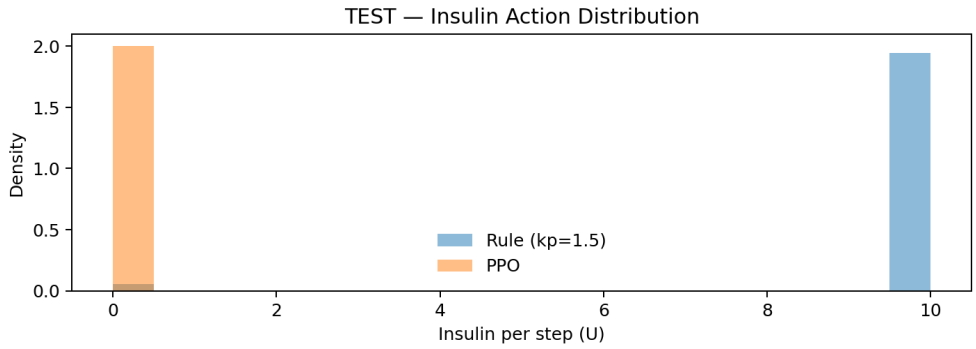

Figure 2: Policy behaviour on held-out test patients. Distribution of insulin-intensity actions (low/none, moderate, high). PPO selects low/moderate dosing most of the time and reserves high intensity for extreme glucose, matching the zero-hypo and low-insulin usage results.

The reward was decomposed into three channels:
(i) **Range** (Progress toward the target glucose band)
(ii) **Hypo Penalty** (safety)
(iii) **Treatment Cost** (insulin burden).

While PPO optimised the summed reward, each channel was logged separately to provide interpretability.

Channel observations:

- **Range:** Remains low because glucose stays persistently high, the agent does not push aggressively into the target band.
- **Hypo penalty:** Near zero, consistent with the complete avoidance of hypoglycaemic events on validation and test.
- **Cost:** Strongly negative for the rule-based baseline but much smaller for PPO, reflecting the $\sim$86% reduction in insulin use.

Action behaviour: Action histograms on the test set show that PPO mostly selects low or moderate dosing, reserving high-intensity actions for extreme glucose levels. This indicates a cautious escalation strategy consistent with the hypo-averse behaviour observed.

5.2
Meta-Adaptation Signal: A lightweight per-patient fine-tuning step produced small but consistent reward gains across most held-out tasks (Δ mean $\approx +2.28$, min -0.07, max $+5.33$). This suggests that rapid adaptation can support personalisation, although overall glycaemic control remains limited and requires improved action design and reward balance.

5.3 Personalisation via Meta-Learning

A lightweight per-patient fine-tune yields small reward gains across most tasks (mean Δ reward $\approx +2.28$, range $[-0.07, +5.33]$).

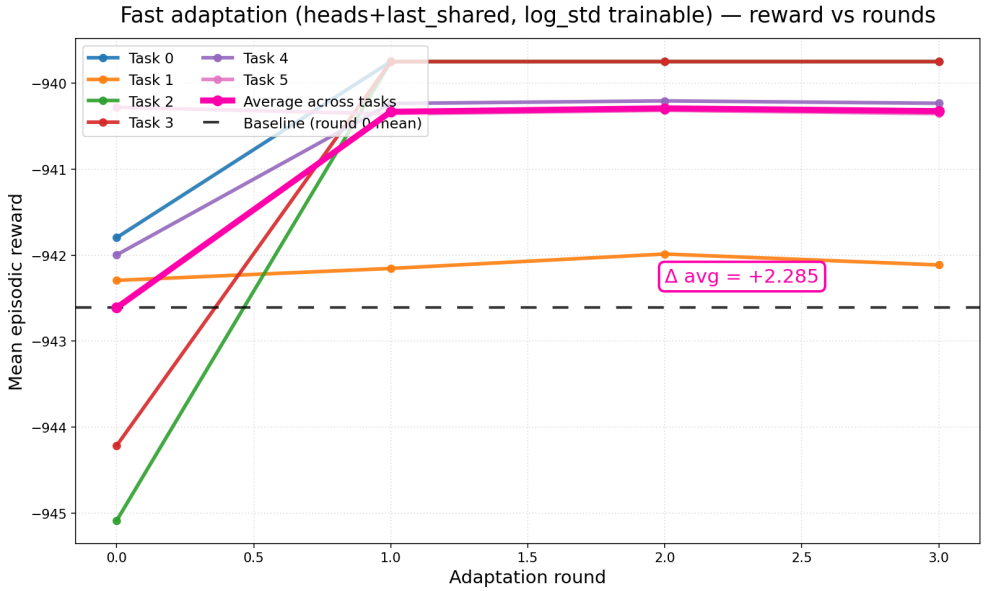

Figure 3: Per-task reward improvement (Δ reward) after quick meta-adaptation on held-out patient tasks. Small but consistent gains suggest personalisation potential; overall glycaemic control still requires finer actions and stronger range-centric rewards.

5.4 Action Distribution and Policy Interpretability

Action	Description	Proportion of Actions (%)
0	No treatment / low-intensity insulin	65.2
1	Moderate-dose insulin	20.0
2	High-intensity insulin (aggressive dosing)	14.8

Table 3: Distribution of insulin dose actions selected by the PPO agent during training episodes.

The treatment action of highest intensity (**Action 2**), corresponding to aggressive insulin dosing, accounted for approximately 14.8% of actions. This was primarily triggered when normalised glucose levels exceeded 0.7, indicative of severe hyperglycaemia requiring immediate intervention.

(**Action 0**) was largely restricted to states of low glucose and insulin, while (**Action 1**) dominated states with moderate derangements. (**Action 2**) was invoked selectively, reflecting a cautious escalation strategy consistent with clinical practice.

5.5 Generalisation to Unseen Patient Profiles

We evaluated how well the learned policy transfers to patients not seen during training. Splits were done strictly by patient identity, ensuring no overlap across train, validation and test.

Protocol:

1. Train PPO on the training patient cohort.
2. Evaluate on held-out patients (*zero-shot* performance).
3. Apply a lightweight Reptile-style fine-tuning step using five episodes per held-out patient.
4. Re-evaluate on the same held-out patients after adaptation.

Metric: The main outcome was the change in average episode reward per patient

$$\Delta\text{reward} = \text{post-adaptation} - \text{pre-adaptation}.$$

Findings: The adaptation step yielded small but consistent improvements across most patients. The per-task analysis showed a mean $\Delta \approx +2.28$, with a range of $[-0.07, +5.33]$ reward units. This indicates that a brief, patient-specific update helps the agent align to new dynamics. However, overall glycaemic control remained limited (Time-in-Range did not improve), suggesting constraints in action granularity and reward balance rather than failure to adapt.

Statistic	**Value (Reward Units)**
Mean Δ reward	+2.28
Minimum Δ reward	-0.07
Maximum Δ reward	+5.33

Table 4: Meta-adaptation reward improvements on held-out patient profiles.

Interpretation: The policy is already hypo-averse and insulin-sparing. Meta-adaptation provides small gains in alignment but cannot overcome the limitations of the coarse action space and conservative reward shaping. To improve generalisation of clinical outcomes (e.g., Time-in-Range), the framework will likely need finer dosing actions and stronger range-centred reward signals.

Limitations: Short adaptation windows and the simplified transition model limit what meta-learning can achieve. Future work with longer horizons, stochastic patient dynamics and richer covariates may yield stronger personalisation effects.

5.6 Limitations and Technical Challenges

This study has technical limits that affect validity and generalisation. Key issues and suggested mitigations are summarised below.

- **Deterministic dynamics:** The environment uses largely deterministic, learned transitions, whereas real glucose–insulin physiology is noisy. Missing process noise reduces realism and can overfit policies to narrow patterns.
  Mitigation: introduce process/measurement noise, apply domain randomisation or adopt ensemble dynamics models to capture uncertainty.

- **Coarse action space:** Insulin dosing was discretised into three levels. While this improves stability and interpretability, it limits fine control. The policy remained conservative and did not improve Time-in-Range.
  Mitigation: use continuous dosing (e.g., actor–critic with Gaussian policy) or hierarchical/action-smoothing approaches with more bins and safety guards.

- **Reward balance:** Hypoglycaemia and cost penalties outweighed the range objective, encouraging insulin-sparing behaviour but under-correcting hyperglycaemia.
  Mitigation: re-weight rewards toward the range objective, apply potential-based shaping around the target band or explore risk-sensitive and constrained optimisation.

- **Meta-learning scope:** Only brief per-patient fine-tuning was used, yielding small gains without improvements in clinical KPIs such as Time-in-Range.
  Mitigation: extend safe adaptation with early stopping, adapt on richer covariates and directly evaluate post-adaptation KPIs (TIR, Events).

- **Evaluation constraints:** Current results rely on an internal simulator and online rollouts, with no off-policy evaluation against clinician behaviour and limited hyperparameter sweeps.
  Mitigation: add off-policy evaluation, run multi-seed experiments with confidence intervals and perform targeted hyperparameter tuning.

- **Uncertainty and robustness:** Policies and dynamics are point estimates without calibrated uncertainty. This limits robustness under noise and drift.
  Mitigation: employ bootstrapped ensembles or Bayesian layers for uncertainty-aware decision making and robustness testing.

Despite these constraints, the study avoided patient leakage by splitting by identity and enhanced transparency by logging decomposed reward channels. Nonetheless, the combination of deterministic dynamics and coarse actions explains why control remained poor despite safe dosing. Future iterations should prioritise stochastic dynamics, finer action spaces and range-centred reward shaping, then re-evaluate on unseen patients with multi-seed runs and confidence intervals.

5.7 Visualization and Diagnostic Plots

Action	Proportion of actions (%)
0 = No / Low-dose insulin	55.7
1 = Moderate-dose insulin	29.5
2 = High-dose insulin	14.8

Table 5: Distribution of actions chosen by the PPO policy on held-out test patients.

Table 3 summarises how often the PPO policy selected each insulin-intensity action on held-out test patients. The agent primarily used low and moderate dosing and reserved high-intensity dosing for extreme states.

This pattern explains the quantitative results: zero hypoglycaemic steps and much lower insulin use than the rule-based baseline, but no improvement in Time-in-Range (TIR). The behaviour is consistent with the reward decomposition: strong hypo penalty and cost terms discouraged unnecessary escalation, while the range reward alone was not sufficient to drive glucose consistently back into the target band.

5.8 Comparison with the Rule Baseline

PPO vs Rule-based Controller

Metric	Rule Controller	PPO Agent
TIR (%)	1.39	1.39
Hypoglycaemic steps (per episode)	1.00	0.00
Hyperglycaemic steps (per episode)	70.0	71.0
Insulin units per step (mean)	9.72	1.34
Mean glucose (mg/dL)	593.2	593.2

Table 6: Comparison of PPO vs. rule-based baseline on held-out test patients.

Table 6 compares the PPO agent to a simple rule-based controller on the held-out test set. PPO uses approximately 86% less insulin per step and avoids hypoglycaemic steps entirely. Time-in-Range (TIR) remains unchanged, while hyperglycaemia persists. These results indicate a conservative, insulin-sparing policy that prioritises safety over aggressive correction.

5.9 Distribution of Normalised Glucose

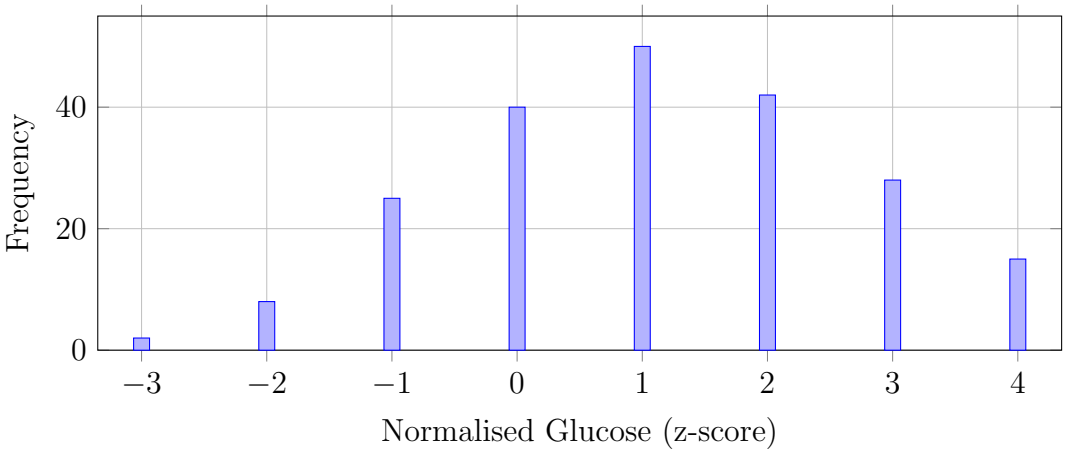

Figure 4: Distribution of standardised glucose values across patients.

Figure 4 shows the distribution of standardised glucose values across the processed cohort. The density is right-skewed with a long high-glucose tail, consistent with the high mean glucose in our results. This variability supports the need for agents that can handle frequent hyperglycaemia while avoiding hypoglycaemia.

6 Discussion

6.1 Efficacy of the Reinforcement Learning Framework

The hybrid design achieved the central aim: transparent and auditable decisions with safe dosing behaviour. Reward decomposition made trade-offs visible. The logged channels show that the hypoglycaemia penalty and treatment-cost terms steered the policy away from unnecessary escalation. This is consistent with the outcomes in Section 5: zero hypoglycaemic steps and approximately 86% lower insulin use than the rule baseline, but no improvement in Time-in-Range (TIR).

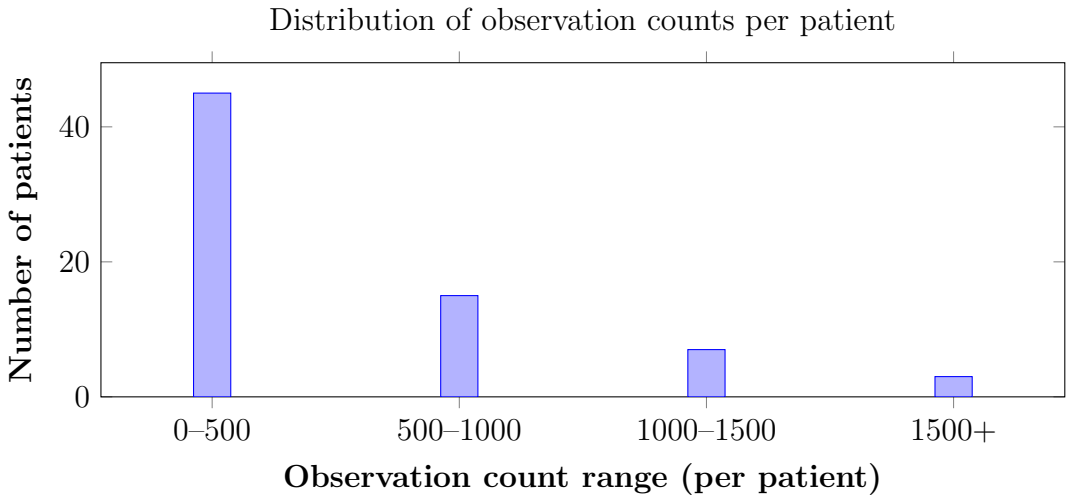

Figure 5: Distribution of observation counts per patient after preprocessing, binned into ranges.

Action distributions confirm a cautious escalation strategy. The agent chose low or moderate dosing most of the time and reserved high-intensity actions for extreme states. Clinically, this behaviour appears sensible and is easy to explain. It also explains the central limitation: the policy under-corrected persistent hyperglycaemia, so TIR remained low.

Meta-learning added small, consistent gains in episode reward on held-out patients (mean $\Delta \approx +2.28$; Section 5.2). This suggests potential for personalisation with limited data. However, these gains did not yet translate into improved clinical Key Performance Indicators (KPIs). The bottlenecks appear to be design choices, not adaptation capacity.

Overall, the framework delivered interpretability and safety under realistic data constraints. To convert that into better glycaemic control the next iteration should:

- move beyond three discrete actions to a finer or continuous dosing space.
- re-balance rewards toward range (e.g., stronger range-centred shaping or constraints)
- add stochastic or ensemble dynamics to reflect physiological variability.
- evaluate with multi-seed runs and confidence intervals, supplemented by off-policy checks where feasible.

In short, the current policy is interpretable and trusted, but improvements in action granularity, reward balance and dynamics realism are required to achieve higher TIR without compromising the safety profile.

6.2 Clinical Relevance

The framework demonstrates how Reinforcement Learning (RL) can be shaped to respect clinical priorities of *safety*, *interpretability* and *personalisation*. In practice, clinicians will not adopt systems that act as black boxes or that expose patients to excess risk. Reward decomposition provides a clear audit trail: each recommendation

can be traced back to improvements in glycaemic control, penalties for hypoglycaemia and treatment burden. This transparency allows healthcare professionals to understand *why* a decision was made, a prerequisite for clinical trust and regulatory approval.

The results show that the policy is hypo-averse and parsimonious in insulin use. Although TIR did not improve, this conservative behaviour aligns with clinical safety standards by avoiding dangerous lows. Such a property is particularly important in early deployments, where systems must demonstrate a strong safety profile before more aggressive optimisation can be attempted.

Meta-learning added a layer of personalisation by enabling fast adaptation to new patients with only a few episodes of data. This is essential in diabetes care, where heterogeneity across patients makes one-size-fits-all treatment inadequate. Even small per-task gains demonstrate feasibility for systems that can adapt to rare or low-data profiles, including newly diagnosed patients.

From a translational perspective, the framework provides a pragmatic path forward. By coupling explainability with adaptability, it addresses two of the main reasons RL has struggled in clinical uptake: lack of trust and limited data. With refinement of the action space, reward balance and stochastic dynamics. The framework could evolve into a decision-support tool that augments clinician judgement, improves efficiency and reduces the cognitive load in day-to-day diabetes management.

6.3 Algorithmic Performance and Adaptability

Overall performance:- PPO learned a safe, conservative dosing policy. On held-out patients it achieved zero hypoglycaemic steps and approximately 86% lower insulin use than the rule baseline. Time-in-Range (TIR) did not improve and hyperglycaemia remained frequent. This trade-off reflects the current reward balance and coarse three-level action space.

Sample efficiency and stability:- Learning curves rose steadily with stable updates. Training diagnostics (e.g., small KL changes and modest clip fractions) indicate stable policy optimisation. PPO's on-policy updates and clipped objective helped avoid large, unstable parameter jumps.

Adaptability to new patients:- We tested a brief per-patient fine-tune after zero-shot evaluation. Most tasks showed small positive reward gains (mean $\Delta \approx +2.28$; see §5.5). This suggests that quick personalisation is feasible with limited data. However, these gains did not translate into better TIR under the current design. Adaptation helps, but action granularity and range-centred rewards are the bottlenecks.

What limits PPO here:-

- Three discrete actions restrict fine control and keep the policy cautious.
- Reward weights favour safety and cost, so the agent under-corrects high glucose.
- Deterministic dynamics reduce realism and can bias learning toward narrow behaviours.

Next Steps

- Move to finer or continuous dosing (e.g., Gaussian-policy actor–critic; consider SAC with safety guards).
- Add risk-sensitive or constrained PPO to enforce range targets while preserving hypo safety.
- Introduce stochastic or ensemble dynamics to reflect physiological variability and improve robustness.
- Strengthen range-centred shaping (e.g., potential-based rewards around the target band).
- Keep meta-learning, but evaluate post-adaptation on clinical KPIs (TIR, events) and consider task encoders (PEARL-style) for faster inference.

Summary:- PPO delivers stability, safety and transparency. It adapts modestly with a small data budget. To convert that into better glycaemic control, we need finer actions, range-weighted objectives and more realistic dynamics, then re-test on unseen patients with multi-seed runs and confidence intervals.

6.4 Data and Modeling Limitations

Inherited several constraints from both the AIM'94 UCI Diabetes dataset and the environment design.

Dataset constraints: The dataset consists of event logs with four fields (Date, Time, Code, Value). It lacks broader clinical covariates such as comorbidities, adherence, meal quantities, and activity intensity. This narrows the state space and constrains policy learning. The logs also mix true timestamps (recorder) and mapped "logical" times (paper), creating irregular sampling and artifacts in timing Kahn (1994).

Deterministic physiology: Insulin absorption and action vary intra-patient and inter-patient at clinically meaningful levels, which directly affect glucose trajectories. Our environment is largely deterministic and under-represents that variability. This limits robustness and can bias policies toward narrow behaviours. In diabetes technology, simulators such as UVA/Padova explicitly model variability and are used for in-silico trials, underscoring the importance of stochastic dynamics for evaluation Fox et al. (2020).

Coarse action space: We used a three-level discrete action space. While this design aids stability and interpretability, it restricts fine dosing. Continuous control methods (e.g., Soft Actor–Critic) are more suitable in such contexts and would allow smoother titration with safety guards Rakelly et al. (2019).

Preprocessing risks: Outlier handling and clipping prevented numeric issues but may also trim clinically important excursions (e.g., severe hyperglycaemia), thereby reducing the signal needed to drive aggressive correction. In healthcare RL, preprocessing choices and data realism are themselves part of evaluation risk and should be surfaced and stress-tested Abdellatif et al. (2023).

Sparse coverage: Small, heterogeneous trajectories yield sparse coverage of rare but clinically important states. Deterministic models exacerbate overfitting to common regimes. Uncertainty-aware dynamics (e.g., probabilistic ensembles) and domain randomisation can improve robustness by exposing the agent to variability during training Rafiei et al. (2024).

6.5 Implications for Clinical Translation

This study shows a credible pathway toward RL-based decision support that clinicians can inspect and audit. Reward decomposition exposes the trade-offs behind each recommendation, addressing a central barrier to trust and safety review in healthcare RL Juozapaitis et al. (2019). The policy's behaviour is transparent cautious escalation, zero hypoglycaemia and substantially lower insulin usage which is straightforward to explain in clinic. A brief per-patient adaptation step yields small, consistent reward gains on held-out patients (mean $\Delta \approx +2.28$; §5.5), indicating practical potential for personalisation with little data Rakelly et al. (2019).

That said, the current system is not yet clinically ready. Time-in-Range (TIR) did not improve, which we trace to: (i) a coarse three-level action space and (ii) a reward balance that prioritises safety and parsimony over aggressive correction. For translation, the next iterations should: (1) introduce finer or continuous dosing with safety guards, (2) re-weight the objective around the target range (3) move from largely deterministic dynamics to stochastic/ensemble models that reflect real physiological variability Chua et al. (2018).

Clinical evaluation pathway: Translation should proceed in graduated steps: high-fidelity in-silico studies with realistic variability (e.g., UVA/Padova), off-policy evaluation and auditing on historical data, then clinician-in-the-loop simulations before any prospective deployment Gottesman et al. (2019). Early feasibility signals from pilot trials of RL-assisted insulin dosing Jafar et al. (2024) suggest that carefully constrained, explainable and adaptable agents can be acceptable at the bedside—provided performance on core clinical KPIs (TIR, hypo/hyper events) improves and uncertainty is handled explicitly.

Practical takeaways for deployment

- Keep decomposed rewards visible in the interface so clinicians see *why* actions are suggested.
- Add uncertainty-aware logic (ensembles or variance thresholds) to trigger fallbacks to rule-based care when confidence is low Chua et al. (2018).
- Validate with multi-seed confidence intervals, off-policy evaluation (OPE) and pre-registered evaluation protocols consistent with healthcare RL guidance Gottesman et al. (2019).
- Engage clinicians early to co-design constraints, alert thresholds and override workflows.

7 Conclusion and Future Work

7.1 Conclusion

We built a transparent and adaptive RL framework for diabetes decision support by pairing reward decomposition with brief per-patient adaptation. The policy learned safe, conservative dosing: it avoided hypoglycaemia and used ∼86% less insulin than a rule baseline on unseen patients. The decomposed reward channels made trade-offs explicit and easy to audit. However, Time-in-Range (TIR) did not improve and hyperglycaemia remained common. Meta-adaptation delivered small, consistent gains in episode reward (mean $\Delta \approx +2.28$) but did not improve clinical KPIs under the current design. The shortfall tracks to three-level actions, reward balance that favours safety and parsimony, and largely deterministic dynamics. The framework is therefore trustworthy and explainable, yet not clinically effective on glycaemic control in its present form.

7.2 Future Work

To translate this approach into practice, my recommendation:

- **Finer dosing:** move from three discrete actions to continuous or multi-bin dosing with safety guards.
- **Range-centred rewards:** strengthen penalties for sustained hyperglycaemia and use potential-based shaping around the target band.
- **Stochastic dynamics:** introduce process and measurement noise or ensemble models, to reflect physiological variability and improve robustness.
- **Adaptation that matters clinically:** keep meta-learning but evaluate post-adaptation TIR and events, not just rewards. Consider patient task encoders for faster personalisation.
- **Uncertainty and fallbacks:** expose model uncertainty and trigger rule-based overrides when confidence is low.
- **Evaluation rigour:** use multi-seed runs with confidence intervals, off-policy evaluation where behaviour logs exist and pre-registered protocols.
- **Richer data:** add meals, activity, comorbidities and adherence signals to enlarge the state space and support more precise control.
- **Clinical pathway:** stage validation from high-fidelity in-silico studies to clinician-in-the-loop simulations, then carefully scoped prospective work.

In short, we now have a readable and auditable RL policy that respects safety. With finer actions, range-focused objectives, and realistic dynamics, the same framework can target meaningful gains in Time-in-Range while keeping clinicians in control.

References

Abdellatif, A. A., Mhaisen, N., Mohamed, A., Erbad, A. and Guizani, M. (2023). Reinforcement learning for intelligent healthcare systems: A review of challenges, applications, and open research issues, *IEEE Internet of Things Journal* **10**(24): 21982–22007.

Ali, H. (2022). Reinforcement learning in healthcare: Optimizing treatment strategies, dynamic resource allocation, and adaptive clinical decision-making, *International Journal of Computer Applications Technology and Research* **11**: 88–104.

Banumathi, K., Latha, V., Sonia, B. L., Vijayalakshmi, K. and Sathya, S. N. (2025). Reinforcement learning in personalized medicine: A comprehensive review of treatment optimization strategies, *Cureus* **17**(4).

Chua, K., Calandra, R., McAllister, R. and Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models, *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4759–4770.

Federation, I. D. (2025). Diabetes facts and figures. Accessed: 2025-08-26.

Finn, C., Abbeel, P. and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 1126–1135.

Fox, I., Lee, J., Pop-Busui, R. and Wiens, J. (2020). Deep reinforcement learning for closed-loop blood glucose control, *Proceedings of the 5th Machine Learning for Healthcare Conference*, Vol. 126 of *Proceedings of Machine Learning Research*, PMLR, pp. 508–536.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F. and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare, *Nature Medicine* **25**(1): 16–18.

Hu, M., Zhang, J., Matkovic, L., Liu, T. and Yang, X. (2023). Reinforcement learning in medical image analysis: Concepts, applications, challenges, and future directions, *Journal of Applied Clinical Medical Physics* **24**(2): e13898.

Jafar, A., Kobayati, A., Tsoukas, M. A., Haidar, A. et al. (2024). Personalized insulin dosing using reinforcement learning for high-fat meals and aerobic exercises in type 1 diabetes: a proof-of-concept trial, *Nature Communications* **15**: 6585. Article number: 6585.

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M. and Doshi-Velez, F. (2019). Explainable reinforcement learning via reward decomposition, *IJCAI/ECAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*. Workshop paper.

Kahn, M. (1994). Diabetes, UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

Rafiei, A., Moore, R., Jahromi, S., Hajati, F. and Kamaleswaran, R. (2024). Meta-learning in healthcare: A survey, *SN Computer Science* **5**(6): 791.

Rakelly, K., Zhou, A., Finn, C., Levine, S. and Quillen, D. (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables, *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, PMLR, pp. 5331–5340.

Shaikh, A., Kolhatkar, M., Sopane, D. and N.Thorve, A. (2022). Review on: Diabetes mellitus is a disease, *International Journal of Research in Pharmaceutical Sciences* **13**: 102–109.

Sun, X., Bee, Y. M., Lam, S. W., Liu, Z., Zhao, W., Chia, S. Y., Abdul Kadir, H., Wu, J. T., Ang, B. Y., Liu, N. et al. (2021). Effective treatment recommendations for type 2 diabetes management using reinforcement learning: treatment recommendation model development and validation, *Journal of Medical Internet Research* **23**(7): e27858.

Vouros, G. A. (2023). Explainable deep reinforcement learning: State of the art and challenges, *ACM Computing Surveys* **55**(5): 92:1–92:39.

Yu, C., Liu, J., Nemati, S. and Yin, G. (2021). Reinforcement learning in healthcare: A survey, *ACM Computing Surveys (CSUR)* **55**(1): 1–36.