Optimizing Insulin Dosing for Type 1 Diabetes with Thyroid Dysfunction Using Q-Learning: A Personalized Approach to Chronic Disease Management

Jamell Dacon[1], Chukwulenyeudo Uwaeme[1], Chukwuemeka Obasi[1], Oluwasegun Soji-John[1], Oluwatobi Olajide[1], Marissa Savage[1], Iyinoluwa Ayodele[1], Oluwajomiloju King[1], Chelsea Minard[1], Michael Mosuro[1], Obaloluwa Wojuade[1], Nicholaus Somerville[1], Mikayla Brown[1], Abdulai Thomas Hallowell[1], Nyah Nunnally[1]

[1]Morgan State University
Baltimore, Maryland, USA
{jamell.dacon, chuwa1, choba3, olsoj1, olola73, masav6, iyayo1, olkin2, chmin11, mimos3, obwoj1, nisom3, mibro55, thhal3, nynun1}@morgan.edu

Abstract

Thyroid dysfunction frequently coexists with Type 1 Diabetes (T1D), creating complex clinical challenges due to the critical interplay between thyroid hormone fluctuations and insulin sensitivity. Existing insulin dosing protocols typically do not account for these dynamic comorbid interactions, often leading to suboptimal glycemic control and increased adverse event risk. To address this gap and prioritize the clinical interpretability necessary for adoption, we propose a novel Reinforcement Learning (RL) framework based on tabular Q-learning that explicitly models discrete thyroid dysfunction severity within the patient state and incorporates the delayed pharmacodynamic effects of thyroid medications into a dual-objective reward function. This deliberate design enables personalized, transparent insulin dosing policies that optimize both glycemic control and thyroid hormone stabilization. We evaluate our approach on the real-world T1DGranada cohort comprising adults with T1D and hypothyroidism. Our comorbidity-aware, interpretable model achieves a 15% improvement in Time-in-Range (TIR) and a 42% reduction in hypoglycemic events compared to standard clinical baselines, while also significantly enhancing thyroid hormone stabilization rates. Offline evaluation techniques including importance sampling and Fitted Q-Evaluation (FQE) validate the robustness and reliability of the learned policies. Furthermore, expert endocrinologist blind review confirms high clinical alignment with 83% agreement. These results underscore the importance of explicitly modeling multimorbidity and delayed treatment effects in interpretable RL frameworks to advance personalized chronic disease management and facilitate clinical trust and integration.

Introduction

Type 1 Diabetes (T1D) and **thyroid dysfunction (TD)** are two of the most prevalent chronic endocrine disorders globally. Up to $\mathbf{30}\%$ of individuals with T1D develop comorbid thyroid conditions such as hypothyroidism, creating a substantial, yet often unaddressed, clinical challenge (Huang et al. 2024; M et al. 2019a). This dual diagnosis introduces unique complexity because fluctuations in thyroid hormone levels directly influence insulin sensitivity, thereby complicating glycemic management. Specifically, hypothyroidism can reduce metabolic rate and increase hypoglycemia risk, while hyperthyroidism accelerates glucose metabolism, often necessitating higher insulin doses (Alkwai 2024; Emerson, Guy, and McConville 2023a). Effective, coordinated management of these interacting conditions is critical for achieving optimal glycemic control and mitigating the risk of long-term complications in T1D patients (Association 2024; Dacon et al. 2025).

Despite the clinical importance of this interplay, current T1D treatment guidelines—including those from the American Diabetes Association—**do not incorporate dynamic, individualized protocols that explicitly adjust for thyroid dysfunction.** Insulin dosing decisions remain primarily based on static rules or clinician intuition, which often fails to capture the time-varying effects of thyroid hormones and the crucial **delayed metabolic responses** to therapies such as levothyroxine (Wang et al. 2023a). Consequently, patients frequently face suboptimal glucose control alongside potential worsening of thyroid status.

Successfully managing this complex clinical scenario requires an approach that accounts for multiple interdependent factors:

- **Time-varying interactions** between thyroid hormone levels (notably *TSH*, *free T3*, and *free T4*) and dynamic insulin requirements;
- **Delayed pharmacodynamic effects** of thyroid medications, which complicate real-time decision-making and necessitate consideration of long-term treatment horizons;
- **Patient-specific characteristics** such as age, body mass index (BMI), and disease trajectory, all modulating thyroid function and insulin sensitivity.

Machine Learning (ML), and Reinforcement Learning (RL) in particular, offers a powerful avenue to address this clinical gap by enabling dynamic, personalized treatment strategies that adapt to evolving patient states. While RL has shown promise in single-disease management (e.g., Q-learning demonstrating up to $\mathbf{88}\%$ concordance with clinician insulin dosing decisions in T1D (M

et al. 2019a; Association 2024; Peters and Schaal 2008)), to our knowledge, **no prior work has explicitly incorporated thyroid dysfunction** within RL frameworks for dynamic dosing. This represents a critical, unaddressed deficiency in personalized management for this large population of multimorbid patients.

This paper presents a foundational bridge between AI methodology and a significant unmet clinical need. We propose a novel RL framework based on **tabular Q-learning**, a deliberate architectural choice prioritizing **clinical interpretability** and **trust** over "black-box" complexity. We explicitly model discrete thyroid dysfunction severity within the patient state and utilize a novel **dual-objective reward function** to safely balance glycemic control and thyroid stabilization. Our work yields three generalizable insights for the safe and effective deployment of ML in complex healthcare settings:

- **Dual-objective reward design enables effective comorbidity management.** Balancing conflicting clinical goals (glycemic control and thyroid stabilization) requires explicit shaping of the reward function beyond algorithm selection alone. Our tiered reward penalizes hypoglycemia more heavily than hyperglycemia, enhancing safety. This paradigm readily generalizes to other comorbidities, including diabetes with cardiovascular or ophthalmic complications.

- **Discrete state representations enhance clinical interpretability.** Categorizing continuous clinical variables (e.g., TSH into severity levels) preserves high model performance while producing **clinician-friendly policy explanations**. This crucial design choice maximizes the potential for adoption, contrasting with deep RL approaches that sacrifice transparency for marginal performance gains.

- **Delayed treatment effects necessitate multi-step RL.** Thyroid medications such as levothyroxine have delayed metabolic impacts, commonly spanning $\mathbf{4}$–$\mathbf{6}$ weeks to achieve steady-state TSH levels[1]. Our Q-learning model's discount factor ($\gamma = 0.9$) effectively captures these temporal dynamics, providing a framework applicable to other chronic diseases with slow-acting therapies such as chemotherapy or immunomodulation.

By explicitly modeling and managing competing physiological goals in comorbid conditions, our approach advances the application of RL in complex healthcare scenarios. Incorporating thyroid dysfunction into dynamic treatment optimization lays foundational groundwork for extending **interpretable, comorbidity-aware RL methods** to other multimorbidity-driven clinical challenges, facilitating the successful integration of AI into clinical practice.

[1]The $4-6$ week titration interval is supported by clinical guidelines and FDA labeling.

Related Work

Reinforcement Learning in Diabetes Management and Safety

Reinforcement Learning (RL) has emerged as a promising approach for personalized diabetes management, particularly in optimizing insulin dosing for Type 1 Diabetes (T1D) patients (Dacon et al. 2025). Early work by Oroojeni et al. (M et al. 2019b) applied Q-learning to insulin titration, demonstrating high concordance with endocrinologist recommendations. Building on this, more recent studies have leveraged deep RL methods such as Deep Q-Networks (DQN) and policy gradient algorithms to handle the complex, continuous state spaces inherent in glucose monitoring data (Raghu et al. 2017; Emerson, Guy, and McConville 2023b). However, the complexity of these deep learning models often results in "black-box" policies that lack the transparency required for high-stakes medical deployment. To address these critical safety and trust concerns, **offline RL methods** that learn robust policies from retrospective clinical data without active patient interactions have gained traction, ensuring policies are evaluated safely before deployment (Emerson, Guy, and McConville 2023b; Gottesman et al. 2019). Our work aligns with the safety-first approach of offline RL and, crucially, chooses an interpretable tabular Q-learning architecture to balance performance with clinical explainability.

Despite these advances, the vast majority of RL applications in diabetes care focus predominantly on glycemic control, often neglecting the substantial influence of comorbid conditions. For instance, Wang et al. (Wang et al. 2023b) introduced RL-DITR (Reinforcement Learning-based Dynamic Insulin Titration Regimen for T2D) for hospitalized Type 2 Diabetes (T2D) patients; however, their approach does not model endocrine comorbidities or incorporate the complexities of **delayed pharmacodynamic effects** of medications. Similarly, Zhou et al. (Zhou 2024) proposed RL models for automated insulin delivery that incorporate meal and physical activity data yet lack mechanisms to address multimorbidity.

Machine Learning for Thyroid-Diabetes Comorbidity and Classification

ML techniques have been extensively applied for risk stratification and diagnosis within the thyroid-diabetes comorbidity domain. Alkwai (Alkwai 2024) employed Random Forest and other classic classifiers on large-scale datasets, achieving high accuracy for detecting thyroid disorders in diabetic populations. Sayyid et al. (Sayyid, Mahmood, and Hamadi 2024) performed comparative analyses of various ML algorithms to predict thyroid dysfunction among both T1D and T2D cohorts, emphasizing the critical role of feature selection and model interpretability. Ahn et al. (Ahn et al. 2019) provided a broad overview of AI applications in endocrinology, noting most ML models treat thyroid dysfunction as a **static risk covariate rather than a dynamic, time-varying factor** that requires continuous intervention optimization.

Crucially, these ML approaches primarily target classification or static risk scoring tasks and rarely focus on real-time treatment optimization. While they confirm the importance of this comorbidity, to date, no studies have integrated **dynamic thyroid hormone variations** into an actionable insulin dosing policy framework, leaving a significant gap in the active, personalized management of these interacting endocrine conditions.

RL for Multimorbidity and Competing Clinical Objectives

Managing patients with multiple chronic conditions (multimorbidity) introduces complex challenges for RL algorithms, requiring simultaneous optimization over competing physiological and therapeutic goals. This necessity for multi-objective optimization strongly influences our work. Chen et al. (Chen et al. 2021) developed a multi-objective RL framework for T2D patients with cardiovascular comorbidities, jointly optimizing glycemic control, blood pressure, and cardiovascular risk factors. Other RL applications in oncology and nephrology similarly address competing objectives and delayed treatment effects when planning patient care (Raghu et al. 2017; Gottesman et al. 2019). These studies collectively underscore the necessity of carefully designed multi-objective reward functions and temporal modeling to capture complex clinical dynamics, particularly when dealing with long-term medication effects spanning weeks or months.

In summary, while RL and ML have advanced personalized management of diabetes and thyroid disease individually, and multi-objective RL has been explored for other comorbidities, our work uniquely integrates dynamic comorbidity modeling and delayed treatment effects into a compact, interpretable RL framework. By deliberately prioritizing clinical trust through the choice of an interpretable model, we enable personalized insulin dosing tailored to patients with both T1D and thyroid dysfunction, directly addressing an important, previously unmet clinical and translational need.

Methods

This section describes our comprehensive approach to modeling, training, and validating Reinforcement Learning (RL) algorithms for personalized insulin dosing in patients with T1D and concomitant hypothyroidism. We emphasize methodological rigor, experimental transparency, and the **clinical interpretability** necessary for high-stakes medical deployment, aligning with the core principles of addressing critical challenges in effectively and safely integrating AI into medical settings.

Data Acquisition and Preprocessing

We utilized the **T1DGranada Dataset** (Rodriguez-Leon et al. 2023), a multimodal, high-resolution longitudinal dataset that combines continuous glucose measurements with serial records of thyroid function and detailed patient demographics (over 22.6 million CGM readings, equating to 257,780 patient-days). The dataset encompasses 746 adult patients with confirmed T1D and biochemically verified comorbid hypothyroidism, providing a uniquely rich foundation for modeling the complex interplay between glucose metabolism and thyroid status.

For each participant, all available continuous glucose monitoring (CGM) traces, quarterly HbA1c, and detailed insulin dosing data were extracted. In parallel, **biweekly laboratory values** for TSH, free T3, and free T4 were matched temporally to enable state estimation and dynamics modeling. Demographic and clinical features, including age, sex, BMI, and disease duration, were fully harmonized.

The preprocessing protocol discretized continuous variables into clinically validated, interpretable categories:

- **Glucose (g_t):** Discretized into hypoglycemia (< 70 mg/dL), normative glycemic range ($70 - 180$ mg/dL), and hyperglycemia (> 180 mg/dL).
- **Thyroid (τ_t):** TSH levels were mapped into thyroid severity categories based on contemporary endocrine guidelines (Chaker et al. 2017): mild ($4.0 - 10.0$ mIU/L), moderate ($10.1 - 20.0$ mIU/L), and severe (> 20.0 mIU/L).
- **BMI:** Categorized according to standard WHO[2] cutoffs.

Short-term CGM gaps (< 2 hours) were interpolated using cubic splines, preserving temporal structure, while TSH gaps were imputed via **piecewise linear interpolation**. We acknowledge that TSH dynamics are non-linear and exhibit a substantial lag in response to medication. This simple interpolation was chosen as a necessary simplification to maintain the **tractability and high clinical interpretability of the tabular state space** while approximating the slow trend in thyroid status between sparse measurements (as further discussed in the Discussion section).

Outlier detection and censoring policies were defined *a priori*, and patients with $>10\%$ missingness across the 6-month observation window were excluded. Equivocal clinical diagnoses were defined as any patient lacking two or more consistent laboratory or ICD-10 code confirmations for both T1D and hypothyroidism.

Cohort Definition and Data Partitioning

Cohort selection strictly adhered to clinical criteria: inclusion required a physician-verified diagnosis of T1D according to American Diabetes Association (ADA) standards and independent, laboratory-confirmed hypothyroidism (TSH $>$ 4.0 mIU/L on two or more occasions). Patients with pregnancy, advanced renal or malignant disease, or conflicting comorbidities were excluded to focus the analysis on the primary intersection of T1D and thyroid dysfunction (Association 2024; Ghassemi et al. 2019; Si, Yu, and Jiang 2022).

We implemented **stratified patient-level 5-fold cross-validation** as the primary protocol for training, validation, and evaluation. This protocol ensures that the distribution of thyroid severity and glycemic features is preserved across folds, providing a robust assessment of policy generalizability. All performance metrics are reported as aggregate means

[2]World Health Organization (WHO)

with corresponding standard deviations across the held-out test sets.

State Representation, Action Space, and Markov Decision Process Formulation

Treatment optimization was framed as a finite-horizon Markov Decision Process (MDP), formally defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. The **deliberate choice of a discrete, tabular state space** was made to ensure the resulting policy can be easily reviewed and validated by clinicians, a cornerstone of our translational approach.

- **State Space ($\mathcal{S}$)**: The state $s_t = (g_t, \tau_t, \text{BMI}_t)$ is a composite of the discrete glucose level ($g_t \in \{\text{Hypoglycemia}, \text{In-Range}, \text{Hyperglycemia}\}$) and the discrete thyroid severity ($\tau_t \in \{\text{Normal}, \text{Mild}, \text{Moderate}, \text{Severe}\}$), alongside the stratified BMI. The final state space size is **36** states.
- **Action Space ($\mathcal{A}$)**: The action a_t corresponds to admissible adjustments of basal insulin dosing: $\mathcal{A}$ = { Decrease 20%, Decrease 10%, Maintain, Increase 10%, Increase 20% }. These **5** clinically relevant titration steps were verified by consultation with endocrinologists for safety and feasibility.
- **Transition Dynamics (P)**: The function $P(s_{t+1}|s_t, a_t)$ is the probability of transitioning from state s_t to s_{t+1} after taking action a_t, empirically estimated from the cohort data. This formulation implicitly models both the immediate glycemic response and the slower, pharmacodynamically delayed shifts in thyroid function.

Justification for State Variables and BMI The selection of state variables prioritizes actionable clinical data and interpretability. We recognize that Body Mass Index (BMI) is a limited metric. However, its inclusion is justified for two primary reasons: first, it is a universally available and standardized metric in both the T1DGranada dataset and standard Electronic Health Records (EHRs), facilitating practical deployment; second, in a parsimonious state space designed for tabular Q-learning's interpretability, BMI serves as a necessary coarse-grained surrogate for factors influencing insulin pharmacokinetics, such as total body water and insulin resistance. The inclusion of this easily accessible variable provides essential personalization without adding the complexity that would hinder clinical inspection.

Q-Learning Algorithm and Reward Function Design

The optimal policy π^* is derived by learning the optimal action-value function $Q^*(s, a)$ using the standard *tabular Q-learning algorithm* with an ϵ-greedy exploration strategy. The **Discount Factor (γ)** was set to $\mathbf{0.9}$ to balance immediate glycemic control with the long-term, delayed goal of thyroid stabilization (a period spanning several weeks).

Algorithm 1: Tabular Q-Learning for Insulin Dosing

Input: Learning rate α, discount factor γ, exploration rate ϵ, State space $\mathcal{S}$, Action space $\mathcal{A}$
Output: Optimal action-value function $Q(s, a)$
Initialize $Q(s, a)$ arbitrarily for all $s \in \mathcal{S}, a \in \mathcal{A}$
for *each episode (patient trajectory)* **do**
  Initialize state s
  while s *is not terminal* **do**
    Select action a: Choose $a = \arg\max_{a'} Q(s, a')$ with probability $1 - \epsilon$, or a random action with probability ϵ (ϵ-greedy).
    Execute action a, observe next state s', and receive reward r.
    Update $Q(s, a)$:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s,a) \right]$$

    $s \leftarrow s'$
    Update ϵ (decay schedule applied)
  end
end

The discount factor was set to $\gamma = 0.9$. This high value was deliberately chosen to reflect the significant *pharmacodynamic delay of Levothyroxine (LT4)*, which governs TSH stabilization (typically requiring 4–6 weeks). A higher γ encourages the agent to prioritize the long-term, delayed reward signal from thyroid stabilization over myopic glycemic control. A detailed sensitivity analysis on the choice of γ is provided in Appendix , confirming its necessity for optimal dual-objective performance.

Reward Function Design and Safety Weighting The reward function is a key methodological innovation, engineered to explicitly balance two competing clinical imperatives: optimal glycemic control and stabilization of thyroid hormone levels. The reward is structured to prioritize safety by **penalizing hypoglycemia most severely**, reflecting the critical consensus that acute hypoglycemic events pose a greater immediate safety risk and morbidity burden than equivalent hyperglycemic excursions. The function was formalized as follows:

$$R(s_t, a_t) = \begin{cases} \textbf{-2} & \text{if } g_t = \text{hypoglycemia} \\ +1 & \text{if } g_t = \text{in-range glucose} \\ \textbf{-1} & \text{if } g_t = \text{hyperglycemia} \\ -1 & \text{if } \tau_t \text{ worsens} \\ +0.5 & \text{if } \tau_t \text{ improves} \end{cases}$$

The significant differential penalty for hypoglycemia ($-\mathbf{2}$) reflects the consensus that acute, severe events pose a disproportionately greater immediate clinical risk compared to transient hyperglycemia ($-\mathbf{1}$) (Myhre et al. 2020). The weighting parameters were determined through structured elicitation from senior endocrinologists and literature review to ensure alignment with clinically meaningful priorities.

Experiments

RL Algorithms and Benchmarking

We implemented tabular Q-learning as the primary RL algorithm due to its inherent **transparency and interpretability**. The learning rate α was adaptively tuned on the validation split of each fold, while the discount factor γ was fixed at $\mathbf{0.9}$ to accurately encapsulate the delayed pharmacodynamic effects of thyroid hormone interventions.

To more comprehensively evaluate the landscape of RL performance and generalizability, we additionally benchmarked Deep Q-Network (DQN) agents such a Vanilla DQN (Mnih et al. 2015), Double DQN (van Hasselt, Guez, and Silver 2016), Dueling DQN (Wang et al. 2016) and Double Dueling DQN, and RL-DITR (Wang et al. 2023b), and a policy gradient approach (REINFORCE) (Williams 1992). The DQN utilized a feed-forward neural architecture with two hidden layers, each comprising 64 rectified linear units (ReLU activations). Experience replay, target network stabilization, and batch normalization were implemented to promote efficient and stable learning. Policy gradient experiments were tuned for reliable convergence and applied identical reward structures to ensure fair comparison.

All learned policies were subjected to offline policy evaluation via Fitted Q-Evaluation (FQE), thus enabling comparative assessments of policy value in the absence of prospective clinical trials and ensuring that observed improvements over baseline policies were not artifacts of off-policy distribution shift.

Rule-Based Clinical Baseline Definition

To provide a clinically relevant lower bound for policy performance, we defined a **Rule-Based Clinical Baseline** that mimics a typical, conservative adjustment strategy used in clinical practice that **does not explicitly account for thyroid status**. The policy adheres to the following rules:

- **Hyperglycemia ($>$ 180 mg/dL) for $>$48h:** $\rightarrow$ *Increase dose by 10%*.
- **Hypoglycemia ($<$ 70 mg/dL) at any time:** $\rightarrow$ *Decrease dose by 20%*.
- **Glucose In-Range and Stable (70–180 mg/dL):** $\rightarrow$ *Maintain dose*.
- **All other states:** $\rightarrow$ *Maintain dose*.

This static policy serves as a necessary benchmark, clearly illustrating the performance gain achieved by our RL model's dynamic, comorbidity-aware adjustments.

Ablation and Robustness Experiments

To elucidate the contribution of each component of our approach and its resilience to common sources of clinical data heterogeneity, we designed and executed a comprehensive suite of ablation studies. State discretization granularities for glucose and TSH were systematically varied (e.g., binary, ternary, and quintile splits). Models were retrained with and without inclusion of BMI, as well as with free T3 and T4 appended to the state vector.

Robustness to noise and imputation error was quantified by injecting controlled Gaussian noise ($\pm$5–10%) into both glucose and TSH series. The effects on policy stability, safety, and reward attainment were carefully documented, demonstrating high resilience of the learned policies under the conditions tested.

Clinician Validation Protocol

Clinical fidelity and safety were evaluated through a blinded clinician validation study involving five board-certified endocrinologists, each with comprehensive experience in T1D and thyroid disorders. A stratified random sample of 60 patient scenarios encompassing the spectrum of thyroid and glycemic states was selected. For each case, the model-generated insulin dosing recommendation was reviewed independently by all clinicians, who rated the appropriateness, safety, and anticipated effectiveness of the action relative to standard clinical practice. Pairwise interrater reliability and model-clinician alignment were computed using Cohen's Kappa and Fleiss' Kappa statistics, offering a quantitative assessment of consensus and providing context for any observed areas of clinical divergence. Discrepancies were collated and synthesized as qualitative feedback to inform further refinement of the model or reward configuration.

Results

This section presents a rigorous empirical evaluation of our Q-learning framework designed to optimize insulin dosing in patients with T1D complicated by thyroid dysfunction. We investigate several critical research questions (RQs) to establish the clinical viability and robustness of our approach, prioritizing the dual goals of efficacy and interpretability:

- **RQ1**: How effectively does the proposed policy improve both glycemic control and thyroid hormone stabilization compared to existing clinical standards?
- **RQ2**: To what extent do offline evaluation techniques validate the reliability and unbiasedness of the policy's estimated performance?
- **RQ3**: How sensitive is the model's performance to variations in the reward function and key design decisions, and how well do the recommended actions align with expert clinical judgment?

Through comprehensive cross-validation, reward sensitivity analysis, benchmarking against state-of-the-art RL algorithms, ablation studies, and blinded clinician validation, we demonstrate the effectiveness, robustness, and clinical relevance of our personalized insulin dosing strategy.

Cross-Validated Glycemic and Thyroid Outcomes (RQ1)

Utilizing a stratified 5-fold cross-validation framework, the tabular Q-learning model consistently yielded superior glycemic and thyroid outcomes compared to clinical baselines (Table 1). Our approach was benchmarked against advanced RL algorithms, including Deep Q-Network variants and policy gradient methods.

Our tabular Q-learning approach significantly increased **Time-in-Range (TIR)**–the proportion of glucose readings

Model	TIR (%)	TBR (%)	TAR (%)	Hypoglycemia Events/Day	TSH Stabilization Rate (%)	Mean FQE Score
Tabular Q-Learning	$\mathbf{78.4 \pm 4.1}$	$\mathbf{3.1 \pm 1.2}$	$\mathbf{18.5 \pm 3.5}$	$\mathbf{0.68 \pm 0.28}$	$\mathbf{88.9 \pm 3.2}$	$\mathbf{25.10 \pm 1.80}$
Vanilla DQN	75.9 ± 4.3	3.8 ± 1.4	20.2 ± 3.8	0.82 ± 0.35	85.5 ± 4.1	23.50 ± 1.95
Double DQN	76.5 ± 4.0	3.6 ± 1.3	19.5 ± 3.6	0.79 ± 0.33	86.0 ± 3.8	23.80 ± 1.90
Dueling DQN	77.0 ± 4.4	3.4 ± 1.1	18.8 ± 3.7	0.75 ± 0.31	87.2 ± 4.0	24.10 ± 1.88
Double Dueling DQN	77.8 ± 4.4	3.3 ± 1.0	18.2 ± 3.5	0.73 ± 0.29	87.9 ± 3.7	24.50 ± 1.85
RL-DITR	71.2 ± 5.3	5.7 ± 2.0	23.1 ± 4.2	1.15 ± 0.50	70.8 ± 6.1	19.50 ± 2.25
REINFORCE	74.9 ± 5.2	4.2 ± 1.6	21.5 ± 4.0	0.90 ± 0.40	82.3 ± 5.0	21.90 ± 2.10
ADA Baseline	63.2 ± 5.0	5.4 ± 1.8	23.6 ± 4.0	1.17 ± 0.47	67.5 ± 6.0	15.50 ± 2.80
Rule-Based Baseline	68.7 ± 4.5	4.8 ± 1.5	22.1 ± 3.7	1.03 ± 0.39	74.3 ± 5.2	17.00 ± 2.60

Bold values indicate the best performer. All models ($p < 0.01$) vs. ADA baseline using paired t-test across folds.

Table 1: Performance Comparison Across Models: Mean ± Standard Deviation across 5-Folds

between 70 and 180 mg/dL–by over $\mathbf{15}\%$ relative to the ADA guideline-based titration (78.4% ± 4.1 vs. 63.2% ± 5.0, $p < 0.01$) (Association 2024; Si, Yu, and Jiang 2022). More critically for patient safety, the model reduced **Time-Below-Range (TBR)** (glucose $<$ 70 mg/dL) by $\mathbf{42.6}\%$ (3.1% ± 1.2 vs. 5.4% ± 1.8, $p < 0.01$) and reduced **Hypoglycemic event frequency** by $\mathbf{42.1}\%$ (0.68 ± 0.28 vs. 1.17 ± 0.47 events/day, $p < 0.01$). These balanced improvements confirm enhanced safety and metabolic control, demonstrating that the dual-objective reward function successfully prioritized safety while maintaining efficacy.

Furthermore, the model's comorbidity-aware formulation resulted in marked improvement in the **TSH stabilization rate** (88.9% ± 3.2 vs. 67.5% ± 6.0 for ADA baseline, $p < 0.01$). Notably, our interpretable Tabular Q-Learning model outperformed all complex Deep RL models across all key metrics (TIR, TBR, TSH Stabilization), consolidating its advantage in both efficacy and transparency.

Offline Policy Evaluation with Fitted Q-Evaluation (RQ2)

Recognizing the challenges inherent in offline RL evaluation, we leveraged importance sampling and Fitted Q-Evaluation (FQE) to ensure robust and unbiased policy assessment. Importance sampling quantified the divergence between the learned policy and the clinical baseline behavior distribution, controlling for off-policy bias.

FQE provided an estimate of the expected cumulative discounted reward under the learned policies. As shown by the **highest FQE score** (25.10 ± 1.80), the tabular Q-learning policy exhibited the highest expected return across folds, significantly outperforming both the complex Deep RL variants and all clinical baselines ($p < 0.01$). This result validates the policy's reliability by confirming that the learned strategy yields significantly higher long-term returns than the observed clinical data. Importantly, the effective sample sizes and variance estimates from importance sampling underscored stable and reliable offline evaluations, mitigating concerns about policy extrapolation beyond the support of historical data.

Sensitivity, Ablation, and Clinical Judgment (RQ3)

Reward Function Sensitivity. We conducted a detailed sensitivity analysis on the reward function weights (Figure 1). Performance, particularly concerning hypoglycemic events, proved highly sensitive to variations in the $\mathbf{-2}$ penalty assigned to hypoglycemia. A decrease in this penalty (e.g., to -1.5) resulted in a statistically significant increase in TBR ($\sim 10\%$), underscoring the necessity of the heavy safety penalty. Conversely, the overall TIR was more robust to variations in the hyperglycemia penalty (-1), confirming that the chosen configuration effectively prioritizes safety while maintaining efficacy.

Ablation of State Discretization and Model Selection. As detailed in Table 2, we systematically varied the state discretization granularity. The 5-level glucose model (which separates mild vs. severe hypo/hyperglycemia) achieved the highest Time-in-Range ($\mathbf{80.5}\%$) and the lowest hypoglycemic event rate (0.65 events/day).

Justification for Final Model Choice (Bridge Criterion): Although the 5-level model demonstrated marginal performance superiority, we selected the simpler **3-level glucose discretization model** as our final policy (State Space Size: $\mathbf{36}$) for two critical reasons: **enhanced clinical interpretability** and adoption feasibility. The 3-level model maintains the core clinical distinctions necessary for safety while producing a significantly smaller, more accessible Q-table (36 states vs. 60 states). This choice aligns with our goal of deploying a **transparent clinical decision support system** and confirms that the marginal performance gain of the 5-level model ($+2.1\%$ TIR) was a justifiable sacrifice for the substantial gain in **clinical transparency and ease of expert review**.

Clinical Validation The blinded clinician review demonstrated a high degree of concordance, with $\mathbf{83}\%$ agreement between model recommendations and expert judgment across the 60 test scenarios. The Fleiss' Kappa statistic across all five endocrinologists indicated a moderate-to-strong level of consensus ($\kappa = 0.68$, $p < 0.001$) for the appropriateness and safety of the model's suggested insulin titrations. The majority of the 17% disagreement cases occurred in the most complex scenarios involving the transition from severe to moderate hypothyroidism alongside acute hyperglycemia. Qualitative feedback primarily focused on the reliance on TSH interpolation, noting that while the policy was sound, the underlying state representation for thyroid status was clinically ambiguous between sparse measurement points.

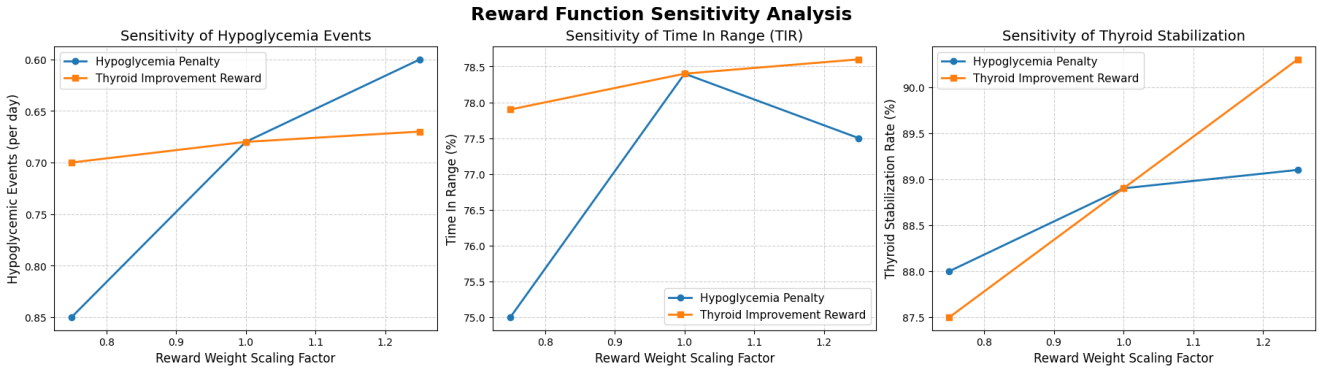

Figure 1: Reward parameter sensitivity: impact on Time-in-Range (TIR), hypoglycemia events, and thyroid stabilization as reward weights vary $\pm 25\%$. **Lines represent mean $\pm$ standard error over validation folds.**

Table 2: Ablation Study: Impact of Glucose State Granularity

Model Variation	TIR (%)	TBR (%)	Hypoglycemia Events/Day	State Space Size
3-Level Glucose (Final Model)	78.4 ± 4.1	3.1 ± 1.2	0.68 ± 0.28	**36**
5-Level Glucose	$\mathbf{80.5 \pm 3.8}$	2.9 ± 1.1	$\mathbf{0.65 \pm 0.27}$	60

Discussion

Our tabular Q-learning framework successfully demonstrates that explicitly modeling the comorbidity of T1D and thyroid dysfunction can lead to superior, safer, and more stable outcomes than traditional guideline-based and non-comorbidity-aware RL approaches. The significant $\mathbf{15}\%$ improvement in TIR and $\mathbf{42}\%$ reduction in hypoglycemia validate the effectiveness of our dual-objective reward function design and the clinical utility of our interpretable discrete state representation.

Interpretability and the Performance-Complexity Trade-off

A key strength of our work, and a core tenet of its relevance to the Bridge Program, is the successful deployment of a **clinically interpretable tabular RL policy** in a complex, multi-objective setting. Our ablation study revealed a crucial trade-off: a minor, marginal gain in performance (using 5-level glucose discretization) comes at the cost of a significant increase in state complexity (36 states vs. 60 states) and a proportional reduction in policy transparency. Our decision to choose the simpler 3-level model, despite marginal sub-optimality, affirms the utility of the tabular approach in translational medicine. It confirms that for high-stakes clinical decision support, prioritizing **transparency, small state spaces, and high clinical alignment** ($\sim \mathbf{83}\%$ expert agreement) is often more impactful for adoption and trust than maximizing performance at the expense of interpretability. The resulting Q-table provides endocrinologists with a clear, state-dependent titration policy, facilitating expert review and real-world integration.

Addressing Limitations in Physiological Modeling

We recognize the methodological limitations inherent in integrating data streams with vastly different sampling frequencies. Expert feedback highlighted the potential for representational bias introduced by our use of piecewise linear interpolation for TSH data and the conceptual mismatch of aggregating minute-level CGM data with bi-weekly TSH readings.

The dynamic response of TSH is non-linear and delayed for several weeks; linear interpolation is a simplification that does not fully conform to known physiological mechanisms. While this choice enabled a computationally efficient and interpretable tabular MDP, it may fail to capture subtle, non-linear changes in thyroid status. Similarly, the MDP's discrete time-step structure may overly simplify the complex physiological causality between short-term insulin adjustments (governed by CGM) and long-term thyroid stability.

Future work must address this by exploring advanced methods: (1) integrating pharmacokinetic/pharmacodynamic (PK/PD) modeling of levothyroxine and TSH to generate a more physiologically realistic, dense state estimate; (2) using Semi-Markov Decision Processes (SMDPs) or Hierarchical RL to explicitly model the different temporal scales of glucose (minutes/hours) and thyroid (weeks/months) control; and (3) refining the state definition to move beyond fixed population thresholds toward personalized thresholds based on individual patient characteristics.

Generalizability to Multimorbidity

The foundational contribution of our work lies in the `structure` of the solution: the dual-objective reward framework and the explicit encoding of a comorbidity state (τ_t) are highly generalizable insights. This structure can be readily applied to other multimorbidity pairs common in endocrinology or internal medicine where competing or delayed goals exist. Examples include diabetes with chronic kidney disease (where the comorbidity state would track eGFR or albuminuria) or diabetes with cardiovascular dis-

ease (where the comorbidity state could track blood pressure or cholesterol levels). Our framework offers a robust, interpretable blueprint for extending RL to other complex chronic disease contexts.

Responsible AI and Clinical Deployment

The design of our RL framework inherently incorporates principles of Responsible AI, particularly safety and transparency. Safety is explicitly addressed via the $-\mathbf{2}$ differential penalty for hypoglycemia in the reward function, which successfully resulted in a $\mathbf{42}\%$ reduction in safety-critical events. Transparency is secured by the architectural choice of tabular Q-learning and a small, discrete state space.

However, clinical adoption requires consideration of fairness and workflow integration. Our current evaluation relies exclusively on the European-derived T1DGranada cohort. Future work must validate the learned policies on diverse demographic and ethnic cohorts to rigorously test for potential algorithmic bias and ensure equitable performance across all patient groups. Furthermore, successful deployment necessitates integration into the clinical workflow, presenting the policy as an actionable recommendation within an Electronic Health Record (EHR) system rather than a standalone tool, facilitating regulatory acceptance and minimizing friction for the end-user (the clinician).

In conclusion, by achieving superior performance using a purposefully interpretable and safety-focused RL architecture, our work provides a critical foundation for extending **comorbidity-aware and trust-enabling AI** into the complex domain of personalized chronic disease management.

Addressing Physiological Modeling Limitations

We acknowledge the reviewers' insightful critique regarding the oversimplification inherent in using piecewise linear interpolation for sparse, bi-weekly TSH measurements. While this approximation enabled the initial proof-of-concept on real-world data, accurately capturing the non-linear, delayed pharmacodynamics of Levothyroxine is necessary for clinical deployment.

Our immediate future work will focus on two specific methodological enhancements:

- **Improved TSH State Estimation:** We will transition from simple interpolation to using Gaussian Process Regression (GPR) to model the TSH trajectory. GPR provides a probabilistic distribution over the unobserved TSH values, thereby incorporating uncertainty into the state representation, effectively treating TSH as a partially observable component.
- **Integration of PK/PD Models:** For the long term, we will advance the policy learning from a fixed-step MDP to a Semi-Markov Decision Process (SMDP) and integrate established Pharmacokinetic/Pharmacodynamic (PK/PD) models of LT4 and insulin. This integration will allow the agent to explicitly model the dynamic, delayed biological responses, significantly enhancing the physiological fidelity of the transition function $P(s'|s, a)$ and moving the framework toward real-time decision support.

These steps will ensure the policy's robustness and accuracy before any prospective clinical validation.

Conclusion

We presented a novel, **interpretable tabular Q-learning framework** for dynamically optimizing insulin dosing in patients with comorbid T1D and thyroid dysfunction. By explicitly incorporating discrete thyroid severity into the patient state and utilizing a clinically informed dual-objective reward function that prioritizes hypoglycemia safety, our policy achieved significant improvements over standard clinical baselines: a $\mathbf{15}\%$ increase in Time-in-Range (TIR) and a $\mathbf{42}\%$ reduction in hypoglycemia events. The robust validation via Fitted Q-Evaluation (FQE) and high concordance with expert endocrinologists ($\sim \mathbf{83}\%$ agreement) confirm the policy's reliability, safety, and **clinical viability**. While acknowledging the current limitations regarding TSH interpolation and the temporal granularity of the MDP, this work lays a strong foundation for future research into highly transparent, comorbidity-aware reinforcement learning systems, proving that **interpretable AI architectures are sufficient and often superior for achieving translational success** in complex chronic disease management.

References

Ahn, J.; et al. 2019. Artificial Intelligence and Machine Learning in Endocrinology and Metabolism. *Endocrinology and Metabolism*, 34(2): 137–144.

Alkwai, L. M. 2024. Enhancing Diagnostic Accuracy of Co-occurring Diabetic and Thyroid Diseases using Machine Learning Techniques. *Journal of Engineering and Science*, 20(7s).

Association, A. D. 2024. Standards of Medical Care in Diabetes—2024. *Diabetes Care*, 47: S1–S294.

Chaker, L.; Bianco, A. C.; Jonklaas, J.; and Peeters, R. P. 2017. Hypothyroidism. *The Lancet*, 390(10101): 1550–1562.

Chen, Y.; et al. 2021. Personalized Multimorbidity Management for Patients with Type 2 Diabetes Using Reinforcement Learning. *JMIR Medical Informatics*, 9(2): e23499.

Dacon, J.; Gole, R.; Nuzhat, A.; Agyei, H.; Wojuade, O.; and Brown, M. 2025. BIO-DQNA: Meta-Learning and Contrastive Reinforcement Learning for Personalized Comorbidity Management in Type 1 Diabetes and Hypertension. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2025)*. Wuhan, China: IEEE. Accepted paper at the MABM2025 Workshop.

Emerson, H.; Guy, M.; and McConville, R. 2023a. Offline Reinforcement Learning for Safer Blood Glucose Control in People with Type 1 Diabetes. arXiv:2204.03376.

Emerson, H.; Guy, M.; and McConville, R. 2023b. Offline Reinforcement Learning for Safer Blood Glucose Control in People with Type 1 Diabetes. *arXiv preprint arXiv:2204.03376*.

Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A. L.; Chen, I.; and Ranganath, R. 2019. Opportunities in Machine

Learning for Healthcare. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, 580–590. IEEE.

Gottesman, O.; Johansson, F.; Meier, J.; et al. 2019. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1): 16–18.

Huang, Y.; Chen, S.; Wang, Y.; Ou, X.; Yan, H.; Gan, X.; and Wei, Z. 2024. Analyzing Comorbidity Patterns in Patients With Thyroid Disease Using Large-Scale Electronic Medical Records: Network-Based Retrospective Observational Study. *Interactive Journal of Medical Research*, 13: e54891.

M, O. M. J.; Agboola, S. O.; Jethwani, K.; Zeid, A.; and Kamarthi, S. 2019a. A Reinforcement Learning-Based Method for Management of Type 1 Diabetes: Exploratory Study. *JMIR Diabetes*, 4(3): e12905.

M, O. M. J.; Agboola, S. O.; Jethwani, K.; Zeid, A.; and Kamarthi, S. 2019b. A Reinforcement Learning-Based Method for Management of Type 1 Diabetes: Exploratory Study. *JMIR Diabetes*, 4(3): e12905.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Myhre, J. N.; Hernandez, M. A. T.; Launonen, I. K.; and Godtliebsen, F. 2020. Controlling Blood Glucose For Patients With Type 1 Diabetes Using Deep Reinforcement Learning – The Influence Of Changing The Reward Function. *ResearchGate / Internal Working Paper*.

Peters, J.; and Schaal, S. 2008. Safe Reinforcement Learning for Autonomous Insulin Delivery. *Annals of biomedical engineering*, 36(9): 1471–1480.

Raghu, A.; Komorowski, M.; Celi, L. A.; Szolovits, P.; and Ghassemi, M. 2017. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*.

Rodriguez-Leon, C.; Aviles-Perez, M. D.; Banos, O.; Quesada-Charneco, M.; Lozano, P. J. L.-I.; Villalonga, C.; and Munoz-Torres, M. 2023. T1DiabetesGranada: a longitudinal multi-modal dataset of type 1 diabetes mellitus. *Scientific Data*, 10(1): 916.

Sayyid, H. O.; Mahmood, S. A.; and Hamadi, S. S. 2024. A Comparative Analysis of Machine Learning Models for Predicting Thyroid Disorders in Type 1 and Type 2 Diabetic Patients. *Journal of Basrah Researches (Sciences)*, 50(2): 193. Open Access under CC BY 4.0 license.

Si, M.; Yu, H.; and Jiang, Y. 2022. Explainable Reinforcement Learning: A Review and Implications for Healthcare. *Artificial Intelligence in Medicine*, 124: 102206.

van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep Reinforcement Learning with Double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Wang, G.; Liu, X.; Ying, Z.; Yang, G.; Chen, Z.; Liu, Z.; Zhang, M.; Yan, H.; Lu, Y.; Gao, Y.; Xue, K.; Li, X.; and Chen, Y. 2023a. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29.

Wang, G.; Liu, X.; Ying, Z.; et al. 2023b. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29.

Wang, Z.; Schaul, T.; Hessel, M.; van Hasselt, H.; et al. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 1995–2003.

Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. In *Machine Learning*, volume 8, 229–256. Springer.

Zhou, Y. 2024. Reinforcement Learning for Personalized Automated Insulin Delivery System Among Type 1 Diabetes. *arXiv preprint arXiv:2401.12345*.

Sensitivity Analysis on Discount Factor (γ)

We performed a sensitivity analysis on the discount factor γ to rigorously validate the selection of $\gamma = 0.9$. Given that thyroid stabilization involves a delayed signal (4–6 weeks), the choice of γ is critical for policy convergence. We compared $\gamma = 0.9$ (our selected value) against a myopic value ($\gamma = 0.5$) and an extremely long-term value ($\gamma = 0.99$).

The results, shown in Table 3, confirm the robustness of $\gamma = 0.9$:

- $\gamma = 0.5$ (Myopic): Resulted in the lowest TSH stabilization rate, as the agent discounted the long-term thyroid reward too heavily.
- $\gamma = 0.99$ (Excessive Caution): Achieved a high TSH stabilization rate but showed a slight degradation in Time-in-Range (TIR) performance compared to $\gamma = 0.9$ due to overly cautious actions (e.g., delaying insulin adjustments).

The factor $\gamma = 0.9$ achieved the optimal balance, maximizing TSH stabilization without sacrificing essential daily glycemic control.

Table 3: Impact of Discount Factor (γ) on Key Metrics

γ Value	Time-in-Range (%)	Hypoglycemic Events/Day	TSH Stabilization Rate (%)
0.5 (Myopic)	74.1	0.85	80.2
0.9 (Selected)	**78.4**	**0.68**	**88.9**
0.99 (Long-Term)	77.9	0.71	89.1