Metaheuristic Algorithms with Applications

www.maa.reapress.com

Metaheur. Algor. Appl. Vol. 1, No. 4 (2024) 417–439.

Paper Type: Original Article

Hybrid Metaheuristic–Reinforcement Learning Framework for Personalized Treatment Policy Optimization

Reem Al-Qahtani[1,*], Arjun Nair[2]

[1] Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; r.qahtani@ksu.edu.sa.

[2] Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India; arjun.nair@iitd.ac.in.

Citation:

Received: 18 February 2024 *Revised: 07 April 2024* *Accepted: 23 July 2024*	Al-Qahtani, R., & Nair, A. (2024). Hybrid metaheuristic–reinforcement learning framework for personalized treatment policy optimization. *Metaheuristic algorithms with applications, 1*(4), 417-439.

Abstract

Personalized medicine increasingly demands dynamic, patient-specific treatment strategies that adapt to evolving clinical states, yet optimizing such strategies remains computationally intractable for conventional approaches. Dynamic Treatment Regimes (DTRs) formalize this challenge as sequential decision-making under uncertainty, but existing Reinforcement Learning (RL) methods for DTR optimization suffer from sample inefficiency, susceptibility to local optima, and safety concerns when trained on limited offline clinical data. This paper introduces the Metaheuristic–Reinforcement Learning (METRO-RL) optimizer, a novel two-stage hybrid framework that synergistically combines Differential Evolution (DE) for global policy search with a Deep Q-Network (DQN) variant for local policy refinement. In *Stage 1*, a modified DE algorithm with patient-state-aware mutation operators explores the policy parameter space globally, identifying promising policy regions while respecting clinical safety constraints. In *Stage 2*, a dueling DQN with Conservative Q-Learning (CQL) regularization fine-tunes the best candidate policies using prioritized experience replay from offline clinical datasets. A novel patient similarity kernel based on demographic features, clinical trajectory alignment via Dynamic Time Warping (DTW), and comorbidity profiles guides both stages, enabling effective knowledge transfer across patient subgroups. We evaluate METRO-RL on two clinically significant applications: Type 2 Diabetes (T2D) management using a simulated 10,000-patient cohort based on UVA/Padova simulator parameters and sepsis treatment in the Intensive Care Unit (ICU) using the MIMIC-IV database (approximately 18,500 sepsis episodes). On T2D management, METRO-RL achieves a composite glycemic control score of 0.782 ± 0.024, representing a 16.2% improvement over standard DQN and an 18.8% improvement over the clinician policy. For sepsis treatment, METRO-RL yields an estimated 90-day mortality of 18.7% ± 2.1%, an 11.4% relative reduction compared to the clinician policy. METRO-RL consistently outperforms standalone DQN, actor-critic, Batch-Constrained Q-Learning (BCQ), CQL, and evolutionary strategy baselines across all Off-Policy Evaluation (OPE) metrics while maintaining the lowest rate of clinically contraindicated actions (0.3%).

Keywords: Dynamic treatment regime, Personalized medicine, Metaheuristic optimization, Reinforcement learning, Clinical decision support, Type 2 diabetes, Sepsis management.

1 | Introduction

The paradigm of personalized medicine has fundamentally transformed the clinical landscape over the past two decades, shifting the standard of care from population-averaged treatment guidelines toward individualized therapeutic strategies tailored to a patient's unique biological, clinical, and behavioral profile [1]. This transformation has been accelerated by the proliferation of Electronic Health Records (EHRs), high-throughput genomic technologies, and wearable sensing devices that collectively generate unprecedented

Corresponding Author: r.qahtani@ksu.edu.sa

https://doi.org/10.48313/maa.v1i4.108

volumes of longitudinal patient data [2]. Despite these data-driven advances, the translation of personalized medicine principles into actionable, optimized treatment policies remains a formidable computational and methodological challenge, particularly for chronic diseases requiring sequential, adaptive therapeutic decisions over extended time horizons.

Dynamic Treatment Regimes (DTRs) provide the formal statistical and decision-theoretic framework for modeling such sequential clinical decision problems [3], [4]. A DTR is a sequence of decision rules, one per treatment stage, that maps patient-specific information accumulated up to each decision point to a recommended treatment action. Optimizing DTRs is intrinsically challenging because clinicians must account for delayed treatment effects, interactions between sequential treatments, patient heterogeneity in response trajectories, and the stochastic nature of disease progression. The goal is to identify an optimal DTR policy that, if followed, would maximize a pre-specified clinical outcome (e.g., survival, quality of life, disease control) across the patient population or within patient subgroups [5].

Classical statistical approaches to DTR estimation include Q-learning [3], [6], [7], marginal structural models with inverse probability weighting [8], g-estimation of structural nested models [4], and Outcome-Weighted Learning (OWL) [9]. While these methods offer strong theoretical guarantees under certain regularity conditions, they often struggle with high-dimensional state spaces, complex nonlinear treatment-response relationships, and the combinatorial explosion of treatment options that characterize modern clinical settings.

The emergence of Deep Reinforcement Learning (DRL) has offered a promising alternative for DTR optimization, leveraging neural network function approximators to handle high-dimensional state representations and learn complex policy mappings. Seminal work by Raghu et al. [10] demonstrated the potential of DQN-based approaches for optimizing intravenous fluid and vasopressor dosing in sepsis management using retrospective ICU data, while subsequent studies have explored RL-based approaches for chemotherapy scheduling [11], anticoagulation dosing [12], and diabetes management [13]. Gottesman et al. [14] provided comprehensive guidelines for evaluating RL-based clinical policies, highlighting the fundamental challenges of Off-Policy Evaluation (OPE) and the risks of deploying RL in safety-critical healthcare settings.

Despite these advances, deep RL methods for DTR optimization face several critical limitations. First, DRL algorithms are notoriously sample-inefficient, requiring large volumes of interaction data that are impractical to obtain in clinical settings where each data point represents a real patient encounter [15]. Second, gradient-based optimization in DRL is susceptible to convergence to poor local optima, particularly in the complex, multi-modal policy landscapes that arise in clinical decision-making with heterogeneous patient populations. Third, offline or batch RL methods [16], [17], which learn from previously collected datasets without additional online interaction, face distributional shift problems where the learned policy may recommend actions poorly represented in the training data, leading to unreliable value estimates and potentially dangerous treatment decisions. Fourth, ensuring safety constraints hard clinical boundaries that must never be violated remains challenging within standard RL frameworks [18].

Metaheuristic optimization algorithms, including evolutionary algorithms, swarm intelligence, and Differential Evolution (DE) [19], offer complementary strengths to RL for policy optimization. These population-based search methods maintain diverse solution sets, exhibit strong global exploration capabilities, and are naturally parallelizable. DE, in particular, has demonstrated effectiveness in high-dimensional continuous optimization with relatively few control parameters [20]. In healthcare contexts, metaheuristic approaches have been applied to treatment scheduling [21], radiotherapy planning [22], and drug dosing optimization [23], though typically for single-decision rather than sequential decision problems.

Recent machine learning work has explored hybrid approaches that combine evolutionary search with gradient-based learning. Neuroevolution methods [24] evolve neural network architectures and weights, Population-Based Training (PBT) [25] optimizes hyperparameters and loss functions during RL training, and evolution strategy–gradient descent hybrids [26], [27] blend population-based exploration with policy gradient refinement. However, these hybrid paradigms have not been systematically adapted for the unique

requirements of clinical DTR optimization, where safety constraints, patient heterogeneity, and offline evaluation requirements introduce domain-specific challenges.

To address these gaps, we propose METRO-RL Optimizer for treatment policy, a novel two-stage hybrid framework for personalized DTR optimization. METRO-RL makes the following five contributions:

I. A hybrid two-stage architecture that combines DE for global policy search with Deep Q-Network (DQN) local refinement, leveraging the complementary strengths of population-based metaheuristic exploration and gradient-based policy optimization.

II. A patient-state-aware mutation operator for DE that adapts mutation intensity based on clinical severity, encouraging broader exploration of aggressive treatments for critically ill patients while maintaining conservative search for stable patients.

III. A novel patient similarity kernel integrating demographic features, DTW–based clinical trajectory alignment, and comorbidity profile similarity, enabling effective knowledge transfer across patient subgroups in both optimization stages.

IV. A comprehensive safety framework combining action masking for hard clinical constraints with Lagrangian relaxation for soft constraint management within a Constrained Markov Decision Process (CMDP) formulation.

V. Extensive evaluation on two clinically significant applications, Type 2 Diabetes (T2D) management and ICU sepsis treatment, demonstrating consistent superiority over eight baseline methods across multiple OPE metrics and clinical outcome measures.

The remainder of this paper is organized as follows. Section 2 reviews related work on DTRs, RL in healthcare, metaheuristics in clinical optimization, and hybrid approaches. Section 3 presents the METRO-RL methodology in detail. Section 4 describes the experimental setup, including datasets, baselines, and evaluation metrics. Section 5 reports and analyzes the experimental results. Section 6 discusses the implications, limitations, and future directions, and Section 7 concludes the paper.

2 | Related Work

2.1 | Dynamic Treatment Regimes

The formal study of DTRs emerged from the causal inference literature in biostatistics, building on Robins's [28] pioneering work on g-computation and structural nested models. Murphy [3] formalized the DTR optimization problem within a potential outcomes framework and introduced A-learning as a semiparametric approach for estimating optimal treatment rules at each stage. Subsequently, Murphy [6] demonstrated the connection between DTR estimation and Q-learning from the RL literature, establishing backward induction with regression-based Q-functions as a practical estimation strategy. This Q-learning approach was further developed by Zhao et al. [29], who studied its theoretical properties, and by Moodie et al. [30], who extended it to accommodate continuous-valued treatments.

The Sequential Multiple Assignment Randomized Trial (SMART) design (Lei et al. [31], Almirall et al. [32]) was developed as the experimental gold standard for comparing embedded DTRs, providing unbiased estimation of regime-specific means under sequential randomization. However, SMART designs are expensive and ethically constrained, motivating the development of methods that can learn DTRs from observational data under the assumption of sequential ignorability (no unmeasured confounders).

OWL [9] and its extensions recast DTR estimation as a weighted classification problem, connecting to support vector machine methods and enabling the use of flexible nonparametric classifiers. Zhang et al. [33] proposed robust OWL methods to handle misspecification, while Zhao et al. [34] extended OWL to multi-stage settings. More recently, tree-based methods [35], [36] have offered interpretable DTR estimation, though they may sacrifice optimality for interpretability.

The Chakraborty and Moodie [5] monograph provides a comprehensive treatment of statistical methods for DTR estimation, covering both parametric and semiparametric approaches, while Tsiatis et al. [37] offer an updated treatment incorporating recent methodological developments. Despite this rich methodological landscape, classical DTR methods have limited capacity to handle the very high-dimensional state spaces and complex nonlinear dynamics characteristic of modern clinical data, motivating the adoption of deep learning-based approaches.

2.2 | Reinforcement Learning in Healthcare

The application of RL to clinical decision-making has accelerated dramatically since the mid-2010s, driven by advances in deep RL and the increasing availability of large-scale EHR datasets. The seminal work of Raghu et al. [10] applied fitted Q-iteration with deep neural networks to optimize vasopressor and intravenous fluid dosing for sepsis patients in the ICU, using retrospective data from the MIMIC-III database. This study demonstrated that RL-derived policies could potentially reduce estimated mortality compared to clinician behavior, though the evaluation relied entirely on off-policy estimation methods. Komorowski et al. [38] extended this line of work with a more clinically nuanced sepsis RL model, incorporating physician expertise into the reward design and performing extensive OPE.

RL has been applied across numerous clinical domains, including: mechanical ventilation weaning [39], where RL policies identified potentially superior weaning strategies; chemotherapy dosing optimization [11], where model-based RL approaches planned drug administration schedules to balance tumor reduction and toxicity; heparin dosing [40], using actor-critic methods for anticoagulation management; and glycemic control in diabetes [13], [41], where RL agents learned insulin dosing policies from continuous glucose monitoring data.

The critical challenge of learning policies from fixed, previously collected clinical datasets has driven significant interest in offline (batch) RL methods. Batch-constrained Q-learning (BCQ) [16] addresses distributional shift by constraining the learned policy to remain close to the behavior policy observed in the data. Conservative Q-learning (CQL) [17] takes an alternative approach, learning a lower bound on Q-values to penalize out-of-distribution actions. Behavior-Regularized Actor-Critic (BRAC) [42] regularizes the actor update to stay close to the behavior policy. These methods have shown improved stability and reliability in offline settings, though they can be overly conservative, potentially failing to discover beneficial treatment strategies that deviate from historical clinician practice.

Safety considerations in RL-based clinical decision support have received increasing attention. Constrained MDP formulations [18], [43] incorporate explicit safety constraints into the optimization objective, while action masking approaches [44] prevent the selection of clinically contraindicated actions. Gottesman et al. [14] provided essential guidelines for the responsible development and evaluation of RL-based treatment policies, emphasizing the limitations of OPE and the dangers of overinterpreting estimated policy improvements.

2.3 | Metaheuristic Optimization in Healthcare

Metaheuristic optimization algorithms have a long history of application in healthcare and biomedical engineering, though primarily for single-stage optimization problems rather than sequential decision-making. Genetic Algorithms (GAs) have been applied to treatment scheduling problems, including radiotherapy fractionation optimization [22], surgical scheduling [45], and chemotherapy regimen selection [21]. Particle Swarm Optimization (PSO) has found applications in medical image segmentation [46], intensity-modulated radiation therapy planning [47], and pharmacokinetic parameter estimation [48].

DE, introduced by Storn and Price [19], has been applied to drug dosing optimization [23], PID controller tuning for glucose regulation systems [49], and model parameter estimation in physiological systems [50]. DE is particularly attractive for policy optimization because it handles continuous parameter spaces efficiently, requires minimal algorithm-specific tuning, and maintains population diversity through its mutation-crossover-selection mechanism [20].

Multi-objective evolutionary optimization has also been applied to clinical problems where multiple competing clinical objectives must be balanced, such as simultaneous minimization of treatment toxicity and maximization of efficacy in cancer treatment [51]. However, applying metaheuristic methods to sequential clinical decision-making under the DTR framework remains largely unexplored.

2.4 | Hybrid Metaheuristic–Machine Learning Approaches

The integration of evolutionary search with gradient-based learning has emerged as a productive research direction in the broader machine learning community. Neuroevolution approaches [24] evolve both the architecture and weights of neural networks, with methods such as NEAT [52] and HyperNEAT [53] demonstrating success on RL benchmarks. Salimans et al. [26] showed that Evolution Strategies (ES) can serve as a scalable alternative to traditional RL algorithms for policy optimization, achieving competitive performance on Atari and MuJoCo tasks while offering superior parallelizability.

PBT [25] combines evolutionary hyperparameter selection with gradient-based weight optimization, allowing simultaneous optimization of the training process itself. Khadka and Tumer [27] proposed Evolutionary Reinforcement Learning (ERL), which maintains a population of actors evolved using genetic operators alongside a gradient-based RL agent, with periodic knowledge transfer between the two components. Pourchot and Sigaud [54] introduced CEM-RL, combining the cross-entropy method with twin delayed deep deterministic policy gradient (TD3) for continuous control. These hybrid approaches consistently demonstrate that combining global population-based search with local gradient-based refinement yields superior performance compared to either approach alone.

Despite this promising body of work, no prior study has systematically developed a hybrid metaheuristic–RL framework specifically designed for clinical DTR optimization. The present work fills this gap by introducing METRO-RL, which adapts the hybrid paradigm to the unique requirements of healthcare applications, including clinical safety constraints, patient heterogeneity, offline evaluation, and the need for interpretable and trustworthy policy recommendations.

3 | Methodology

3.1 | Problem Formulation

We formulate the personalized treatment policy optimization problem as a finite-horizon Markov Decision Process (MDP) defined by the tuple $M = (S, A, P, R, \gamma, T)$, where S denotes the state space representing patient clinical features, A denotes the discrete action space of available treatment options, $P: S \times A \times S \rightarrow [0, 1]$ represents the state transition dynamics capturing disease progression under treatment, $R: S \times A \times S \rightarrow R$ is the reward function encoding clinical outcomes, $\gamma \in [0, 1]$ is the discount factor, and T is the treatment horizon (number of decision stages).

A treatment policy $\pi: S \rightarrow A$ maps the current patient state to a treatment action. The objective is to find the optimal policy π^* that maximizes the expected cumulative discounted reward:

$$\pi * = \arg\max\pi E\pi [\Sigma t = 0T \gamma t R(st, at, st + 1)], \quad (1)$$

The policy is parameterized by a neural network with weights θ, yielding $\pi_\theta(s) = \arg\max_a Q_\theta(s, a)$, where Q_θ is the action-value function approximated by the network.

T2D-MDP

The state space for T2D management is 12-dimensional, comprising: Glycated hemoglobin (HbA1c, %), fasting plasma glucose (FPG, mg/dL), body mass index (BMI, kg/m²), estimated glomerular filtration rate (eGFR, mL/min/1.73m²), age (years), diabetes duration (years), systolic blood pressure (mmHg), low-density lipoprotein cholesterol (LDL, mg/dL), prior medication history (categorical), cumulative hypoglycemia count, exercise level (categorical), and dietary compliance score. The action space consists of 20 discrete treatment

options formed by the combination of 5 metformin dose levels (0, 500, 1000, 1500, 2000 mg/day) and 4 insulin dose levels (0, 10, 20, 40 units/day). The treatment horizon is T = 12 monthly decision points.

Sepsis MDP

The state space for ICU sepsis management is 46-dimensional, comprising demographics (age, sex, weight), vital signs (mean arterial pressure, heart rate, SpO2, temperature, respiratory rate), laboratory values (lactate, creatinine, blood urea nitrogen, white blood cell count, platelet count, bilirubin, international normalized ratio, bicarbonate, partial pressure of oxygen), clinical scores (SOFA score, Glasgow Coma Scale), ventilation status, and treatment history features. The action space consists of 25 discrete actions formed by the combination of 5 vasopressor dose bins and 5 intravenous fluid dose bins. The treatment horizon spans a 48-hour ICU stay discretized into 12 four-hour intervals (T = 12).

3.2 | Metaheuristic–Reinforcement Learning Architecture

The METRO-RL framework employs a two-stage architecture that exploits the complementary strengths of metaheuristic global search and deep RL local refinement. *Stage 1* uses a modified DE algorithm to search the policy parameter space globally, identifying promising regions that are subsequently refined by a modified DQN in *Stage 2*. This two-stage approach is motivated by the observation that gradient-based RL methods are highly sensitive to initialization [55], and that providing a strong initial parameter configuration from global search substantially improves the quality and stability of the final policy.

Stage 1 (Global search via modified DE). We maintain a population of N = 100 candidate policies, each encoded as the weight vector θ_i of a Small Multilayer Perceptron (MLP) with 2 hidden layers of 64 units each and ReLU activations. The network maps the patient state $s \in S$ to action-value estimates $Q_\theta(s, a)$ for all actions $a \in A$. The DE operates through iterative mutation, crossover, and selection to evolve this population toward higher-fitness policy parameters.

3.2.1 | Patient-state-aware mutation

We introduce a novel modification to the standard DE/rand/1 mutation strategy. For each target vector θ_i, the mutant vector is generated as

$$\mathrm{vi} = \theta r1 + F(\Psi) \cdot (\theta r2 - \theta r3), \quad (2)$$

where r1, r2, r3 are distinct randomly selected indices, and $F(\Psi)$ is the adaptive mutation factor that depends on the clinical severity profile Ψ of states encountered during policy evaluation. Specifically:

$$F(\Psi) = \mathrm{Fbase} + \Delta F \cdot \sigma\big(\mathrm{severity}(\Psi)\big), \quad (3)$$

where $F_{base} = 0.3$ is the minimum mutation factor, $\Delta F = 0.7$ is the severity-dependent scaling range, $\sigma(\cdot)$ is the sigmoid function, and severity (Ψ) is computed as the mean normalized clinical acuity score across states visited during policy rollout. This mechanism increases exploration intensity for policies that encounter clinically severe states, encouraging the discovery of aggressive but potentially life-saving treatment strategies for critically ill patients, while maintaining a more conservative search for policies operating in stable clinical contexts.

3.2.2 | Crossover with clinical constraint masking

We employ binomial crossover to generate trial vectors u_i from the target θ_i and mutant v_i. The crossover rate CR = 0.9 determines the probability that each parameter is inherited from the mutant. Crucially, after crossover, we apply a clinical constraint mask that projects any parameter configurations violating predefined safety bounds back to the feasible region. This masking ensures that all candidate policies in the population satisfy hard clinical constraints (e.g., action weights that would recommend insulin during hypoglycemia are zeroed).

3.2.3 | Fitness evaluation

The fitness of each candidate policy $\pi_{\theta i}$ is evaluated as the expected cumulative reward estimated via Monte Carlo rollouts on the patient simulator (for T2D) or Weighted Importance Sampling (WIS) on the offline clinical dataset (for sepsis):

$$\text{Fitness}(\theta i) = \left(\frac{1}{M}\right) \Sigma j = 1M \, \Sigma t = 0T \, \gamma t \, R(stj, \pi\theta i(stj), st + 1j), \quad (4)$$

where M is the number of evaluation trajectories. For the offline setting, we replace the Monte Carlo estimate with the WIS estimator:

$$VWIS(\pi\theta i) \; = \; (\Sigma j = 1n \, wj \, Gj \,) \, / \, (\Sigma j = 1n \, wj \,), \quad (5)$$

where $w_j = \Pi_t \, \pi_{\theta i}(a_t^j \,|\, s_t^j) \; / \; \pi_b(a_t^j \,|\, s_t^j)$ is the importance weight, and G_j is the observed return for trajectory j under the behavior policy π_b.

Algorithm 1. METRO-RL Stage 1 DE-based global policy search.

Algorithm 1: METRO-RL Stage 1 DE-based Global Policy Search
Input: Population size N, generations G_max, base mutation factor F_base, crossover rate CR, patient dataset D, simulator Env (if available) Output: Best policy parameters θ* 1: Initialize population {θ_1, θ_2, ..., θ_N} with Xavier initialization 2: Compute patient clusters C_1, ..., C_K via spectral clustering on kernel K 3: for g = 1 to G_max do 4: for i = 1 to N do 5: Evaluate π_{θ_i} on patient trajectories; compute severity(Ψ_i) 6: F_i ← F_base + ΔF · sigmoid(severity(Ψ_i)) 7: Select distinct r1, r2, r3 ≠ i uniformly at random 8: v_i ← θ_{r1} + F_i · (θ_{r2} - θ_{r3}) // Mutation 9: for each parameter dimension d do 10: if rand() < CR or d = d_rand then 11: u_i[d] ← v_i[d] 12: else 13: u_i[d] ← θ_i[d] 14: end if 15: end for 16: Apply clinical constraint mask to u_i // Safety projection 17: Apply patient similarity transfer: 18: For top-performing policies in similar clusters, 19: blend parameters: u_i ← (1-λ)u_i + λθ_{best,cluster} 20: if Fitness(u_i) ≥ Fitness(θ_i) then 21: θ_i ← u_i // Selection 22: end if 23: end for 24: Record best fitness and update convergence history 25: end for 26: θ* ← arg max_{θ_i} Fitness(θ_i) 27: return θ*

Stage 2 (Local refinement via modified DQN). The second stage initializes a DQN with the best policy weights θ* obtained from *Stage 1* and performs gradient-based local refinement using the offline clinical dataset. Our DQN variant incorporates three key modifications tailored for clinical policy optimization:

3.2.4 | Dueling architecture

Following Wang et al. [56], we decompose the Q-function into a state-value stream V(s) and an advantage stream A(s, a), allowing the network to independently assess how good a patient state is and the relative benefit of each treatment action. This decomposition is particularly valuable in clinical settings where many states have similar overall value regardless of the treatment chosen (e.g., stable patients who will improve under any reasonable treatment).

3.2.5 | Conservative Q-learning penalty

To mitigate overestimation of Q-values for actions underrepresented in the offline dataset, we incorporate a CQL-style penalty [17] into the training loss:

$$L(\theta) = \; LDQN(\theta) + \; \alpha CQL \; \cdot \; Es\sim D \left[\log \Sigma a \, \exp\big(Q\theta(s, a)\big) - \; Ea\sim\pi b[Q\theta(s, a)] \right], \quad (6)$$

where $\alpha_{CQL} = 1.0$ controls the conservatism strength, and L_{DQN} is the standard temporal difference loss with prioritized experience replay [57].

3.2.6 | Safety layer

An action masking layer is applied to the Q-network output before action selection, setting Q-values of clinically contraindicated actions to $-\infty$ based on the current patient state. Contraindicated actions are defined by clinical domain knowledge: for T2D, insulin is masked when the patient is in a hypoglycemic state (FPG < 70 mg/dL); for sepsis, high-dose vasopressors are masked when MAP exceeds 80 mmHg, and excessive fluid boluses are masked when fluid balance is strongly positive.

Algorithm 2. METRO-RL Stage 2 DQN local refinement.

Algorithm 2: METRO-RL Stage 2 DQN Local Refinement
Input: Best DE policy parameters θ*, offline dataset D, patient similarity kernel K, safety constraints C Output: Refined policy parameters θ_refined 1: Initialize Q-network Q_θ with weights θ* 2: Initialize target network Q_θ' ← Q_θ 3: Construct patient-specific replay buffers using kernel K: For patient p, prioritize transitions from similar patients 4: for step = 1 to N_steps do 5: Sample mini-batch B from prioritized replay buffer 6: for each transition (s, a, r, s') in B do 7: Apply safety mask M(s') to Q_θ'(s', ·) 8: y ← r + γ · max_{a' ∉ masked} Q_θ'(s', a') 9: Compute TD error: δ = y - Q_θ(s, a) 10: end for 11: Compute L_DQN = mean(|δ|^2) with importance sampling weights 12: Compute L_CQL = α · [log(Σ_a exp(Q_θ(s,a))) - E_{a~π_b}[Q_θ(s,a)]] 13: L_total = L_DQN + L_CQL 14: Update θ via Adam optimizer: θ ← θ - η ∇_θ L_total 15: if step mod target_update_freq = 0 then 16: Q_θ' ← Q_θ // Hard target update 17: end if 18: end for 19: θ_refined ← θ 20: return θ_refined

3.3 | Patient Similarity Kernel

A central innovation of METRO-RL is a patient similarity kernel that guides knowledge transfer across patient subgroups in both optimization stages. For patients p_i and p_j, the composite similarity kernel is defined as:

$$K(pi, pj) = wd \cdot Kdemo(pi, pj) + wc \cdot Kclin(pi, pj) + wm \cdot Kmorb(pi, pj), \quad (7)$$

where $w_d + w_c + w_m = 1$ are learned kernel weights (default: $w_d = 0.2$, $w_c = 0.5$, $w_m = 0.3$).

Demographic kernel K_{demo} computes similarity over static patient features (age, sex, BMI, baseline eGFR) using a Gaussian Radial Basis Function (RBF):

$$Kdemo(pi, pj) = \exp(-||xidemo - xjdemo||^2 / (2\sigma d^2)). \quad (8)$$

Clinical trajectory kernel K_{clin} uses Dynamic Time Warping (DTW) to measure the alignment cost between multivariate clinical time series (vital signs, lab trajectories):

$$Kclin(pi, pj) = \exp(-DTW(Ti, Tj) / \sigma c), \quad (9)$$

where T_i and T_j are the clinical time series of the two patients, and σ_c is a bandwidth parameter estimated from the data.

Comorbidity kernel K_{morb} computes the Jaccard similarity coefficient over binary comorbidity profiles encoded from ICD-10 diagnosis codes:

$$Kmorb(pi, pj) = |ICDi \cap ICDj| / |ICDi \cup ICDj|, \quad (10)$$

The composite kernel is used for two purposes. In *Stage 1*, spectral clustering on the kernel matrix partitions patients into K = 10 subgroups, and successful policy parameters from high-performing individuals in one cluster are probabilistically transferred to similar clusters during the DE mutation step (lines 17–19 in *Algorithm 1*). In *Stage 2*, the kernel guides the construction of patient-specific replay buffers. When training the DQN on a particular patient subgroup, transitions from patients with high kernel similarity are prioritized in the experience replay buffer, effectively creating a personalized training dataset that enriches the learning signal for underrepresented patient types.

3.4|Reward Function Design

The design of clinically meaningful and computationally tractable reward functions is critical for the success of any RL-based treatment policy. We design domain-specific reward functions for each clinical application that balance treatment efficacy, safety, and patient burden.

3.4.1|T2D reward function

$$\mathrm{RT2D} = -\alpha\,|\mathrm{HbA1c} - \mathrm{target}| - \beta \cdot \mathrm{hypo\_events} - \gamma r \cdot \mathrm{adverse\_effects} + \delta \cdot \mathrm{adherence\_proxy}, \quad (11)$$

where $\alpha = 1.0$ penalizes deviation from the target HbA1c (7.0% for most patients, individualized based on age and comorbidities), $\beta = 2.0$ penalizes hypoglycemic events (FPG < 70 mg/dL), $\gamma_r = 0.5$ penalizes adverse effects (gastrointestinal symptoms, weight gain), and $\delta = 0.3$ provides a small bonus for treatment regimens with higher expected adherence (fewer daily doses, simpler combinations). The target HbA1c is individualized: 7.0% for patients aged <65 without significant comorbidities, 7.5% for elderly patients or those with cardiovascular disease, and 8.0% for patients with limited life expectancy or extensive comorbidity burden.

3.4.2|Sepsis reward function

$$\mathrm{RSepsis} = \lambda 1 \cdot \mathrm{survival90d} + \lambda 2 \cdot (-\Delta \mathrm{SOFA}) + \lambda 3 \cdot (-\mathrm{ICU\_LOS}) - \lambda 4 \cdot \mathrm{excess\_treatment}, \quad (12)$$

where $\lambda_1 = 15.0$ assigns a large terminal reward for 90-day survival, $\lambda_2 = 1.0$ rewards SOFA score reduction (organ function improvement), $\lambda_3 = 0.1$ provides a small per-step penalty for ICU length of stay, and $\lambda_4 = 0.5$ penalizes excessive treatment (vasopressor doses above the 90th percentile of clinician behavior or fluid volumes exceeding 30 mL/kg/6h). The terminal survival reward is assigned only at episode end (discharge or death), while intermediate rewards for SOFA change, length of stay, and excess treatment provide dense learning signals.

To mitigate reward hacking the tendency of RL agents to exploit reward function loopholes rather than achieving genuine clinical improvement we employ several safeguards: 1) reward components are normalized to comparable scales before combination, 2) the excess treatment penalty prevents policies that achieve marginal improvements through aggressive overtreatment, and 3) the OPE battery (Section 3.5) provides independent validation of policy quality beyond the training reward.

3.5|Offline Policy Evaluation

Since prospective clinical deployment is infeasible without prior regulatory approval and clinical trial validation, we employ a comprehensive battery of OPE methods to estimate the value of candidate policies using the historical clinical dataset. Our OPE battery includes four complementary estimators:

3.5.1|Weighted Importance Sampling

Precup et al. [58] provide a consistent, self-normalized estimator of the policy value that reduces variance compared to ordinary importance sampling at the cost of introducing a small bias. The WIS estimate is given by *Eq. (5)*.

Per-Decision Importance Sampling (PDIS) (Precup et al. [58]) improves upon trajectory-level importance sampling by applying the importance correction at each decision step, substantially reducing variance in long-horizon settings:

$$\mathrm{VPDIS}(\pi) = (1/n)\,\Sigma j = 1n\,\Sigma t = 0T\,\gamma t\,(\Pi k = 0t\,\pi(akj|skj) / \pi b(akj|skj))\,Rtj. \quad (13)$$

Fitted Q-evaluation (FQE) (Le et al. [59]) trains a separate Q-network to evaluate the target policy by iterating the Bellman evaluation equation, providing an estimate that does not rely on importance weights and is thus more stable when the target policy differs substantially from the behavior policy.

Doubly Robust (DR) estimator (Jiang and Li, [60]) combines importance sampling with a Q-function baseline to achieve lower variance than pure IS while remaining unbiased when either the importance weights or the Q-function model is correctly specified. Agreement across these four estimators provides stronger confidence in the estimated policy value than any single estimator alone.

3.6 | Safety Constraints

We formulate clinical safety within the CMDP framework. Let $C_k(s, a)$ denote the k-th constraint cost function, with corresponding threshold d_k. The constrained optimization problem is:

$$\max\pi \, E\pi[\Sigma \, \gamma t \, Rt] \;\; \text{subject to} \;\; E\pi[\Sigma \, \gamma t \, Ck(st, at)] \leq dk, \;\; \forall \, k. \tag{14}$$

We address constraints at two levels. Hard constraints represent absolute clinical contraindications and are enforced via action masking, which sets $Q(s, a) = -\infty$ for contraindicated actions in both the DE fitness evaluation and DQN action selection. The action-masking mechanism ensures zero probability of selecting a dangerous action at any point during optimization or deployment. Soft constraints represent relative clinical warnings and are handled via Lagrangian relaxation, augmenting the reward with penalty terms:

$$Raug(s, a, s') = R(s, a, s') - \Sigma k \, \mu k \cdot Ck(s, a), \tag{15}$$

where $\mu_k \geq 0$ are Lagrange multipliers updated via dual gradient ascent during *Stage 2* training to adaptively tighten or relax soft constraints based on observed constraint violations.

4. | Experimental Setup

4.1 | Type 2 Diabetes Dataset

The T2D dataset comprises a synthetic cohort of 10,000 virtual patients generated using a disease progression model inspired by the UVA/Padova glucose-insulin simulator [61], [62]. While the original UVA/Padova simulator focuses on Type 1 Diabetes with minute-level glucose dynamics, we extend the approach to model T2D progression at a monthly granularity, incorporating long-term disease evolution features including progressive beta-cell function decline, insulin resistance dynamics, weight changes, and renal function trajectories. Patient parameters are sampled from distributions calibrated to the United Kingdom Prospective Diabetes Study (UKPDS) cohort demographics [63]: age uniformly distributed between 40 and 75 years, BMI between 25 and 40 kg/m², baseline HbA1c between 7.0% and 10.5%, and baseline eGFR between 45 and 120 mL/min/1.73m². Comorbidities (cardiovascular disease, neuropathy, nephropathy, retinopathy) are assigned probabilistically based on age, diabetes duration, and baseline HbA1c, using prevalence rates from the Diabetes Control and Complications Trial and UKPDS studies.

The treatment horizon is 12 months with monthly decision points. At each decision point, the clinician (or policy) selects from 20 discrete treatment options: 5 metformin dose levels (0, 500, 1000, 1500, 2000 mg/day) combined with 4 insulin dose levels (0, 10, 20, 40 units/day). The simulator propagates the patient state forward by one month, incorporating stochastic noise to model individual variability in drug response, adherence fluctuations, and lifestyle factors. The dataset is split into 7,000 patients for training, 1,500 for validation, and 1,500 for testing.

4.2 | Sepsis Dataset (MIMIC-IV)

The sepsis cohort is extracted from the MIMIC-IV database [64], a freely available critical care database comprising de-identified health-related data associated with patients admitted to the Intensive Care Units (ICUs) of the Beth Israel Deaconess Medical Center. We identify sepsis episodes using the Sepsis-3 criteria [65]: suspected infection (defined by concurrent antibiotic administration and body fluid cultures) combined

with an acute change in SOFA score of ≥2 points. After applying inclusion criteria (age ≥18, ICU stay ≥24 hours, sufficient data completeness ≥80% of features recorded) and exclusion criteria (readmissions, transfers from other ICUs, patients with do-not-resuscitate orders at admission), we obtain approximately 18,500 sepsis episodes.

The 48-hour treatment window following sepsis onset is discretized into 12 four-hour intervals. The 46-dimensional state vector is computed at each interval using the most recent available measurements, with forward-filling for missing values and population-median imputation for features never recorded. States are normalized to zero mean and unit variance using training set statistics. Treatment actions (vasopressor dose and IV fluid volume within each 4-hour interval) are discretized into 5 bins each using quintiles of the clinician dose distribution, yielding 25 discrete actions. The primary outcome is 90-day all-cause mortality, determined from hospital records and the Social Security Death Index. The dataset is split by admission year: 2008–2017 for training (approximately 14,800 episodes), 2018–2019 for validation (approximately 1,850 episodes), and 2020–2022 for testing (approximately 1,850 episodes).

4.3 | Compared Methods

We compare METRO-RL against eight baseline methods spanning pure RL, offline RL, and pure evolutionary approaches:

I. Standard DQN (Mnih et al. [66]): DQN with experience replay and target network, initialized with random weights.

II. Dueling DQN (Wang et al. [56]): DQN with value-advantage stream decomposition, without CQL penalty or DE initialization.

III. BCQ (Fujimoto et al. [16]): BCQ with generative model for action proposal, designed for offline RL.

IV. CQL (Kumar et al. [17]): CQL with pessimistic value estimation for out-of-distribution actions.

V. A2C (Mnih et al. [67]): advantage actor-critic with offline training via importance-weighted policy gradient.

VI. OpenAI ES (Salimans et al. [26]): evolution strategy with antithetic sampling and fitness shaping, same network architecture as METRO-RL.

VII. CMA-ES (Hansen [68]): covariance matrix adaptation evolution strategy for direct policy search, adapted for neural network policy parameterization.

VIII. Clinician policy: the observed behavior policy estimated from the training data via behavior cloning (multinomial logistic regression).

Additionally, a random policy baseline is included for reference. All neural network-based methods use the same architecture (2 hidden layers, 64 units, ReLU) for fair comparison. All methods (except the clinician and random baselines) are trained on the same training data and evaluated on the same held-out test set.

4.4 | Evaluation Metrics

We evaluate all methods using both OPE metrics and domain-specific clinical outcome metrics.

OPE metrics: WIS estimated value, FQE estimated value, and DR estimated value, each computed on the test set with 95% confidence intervals via bootstrap resampling (1,000 replicates).

T2D clinical metrics: 1) mean HbA1c reduction from baseline over the 12-month horizon, 2) hypoglycemia rate (percentage of monthly time steps with FPG < 70 mg/dL), and 3) composite glycemic control score, defined as: Score = 0.5 · norm(HbA1c_reduction) + 0.3· (1 − norm(hypo_rate)) + 0.2· norm(adherence), where norm denotes min-max normalization across methods.

Sepsis (Clinical metrics): 1) estimated 90-day mortality rate under the learned policy (via WIS), 2) estimated SOFA score reduction (mean change from admission to 48h), 3) estimated vasopressor-free hours within the 48-hour window, and 4) estimated ICU length of stay.

Safety metrics: 1) rate of contraindicated actions (percentage of state-action pairs violating hard clinical constraints), and 2) treatment volatility index, defined as the mean absolute change in treatment intensity between consecutive decision points, normalized by the clinician policy's volatility.

4.5 | Hyperparameters and Implementation Details

Table 1 summarizes the hyperparameters for METRO-RL and all baselines. All experiments are repeated over 30 independent runs with different random seeds to assess statistical reliability.

Table 1. Hyperparameters and implementation settings of METRO-RL and baseline methods.

Component	Parameter	Value
Stage 1 (DE)	Population size N	100
Generations G_{max}	200	
Base mutation factor F_{base}	0.3	
Mutation range ΔF	0.7	
Crossover rate CR	0.9	
Stage 2 (DQN)	Learning rate	1×10^{-4}
Replay buffer size	100,000	
Target network update	Every 1,000 steps	
Training steps N_{steps}	50,000	
CQL penalty α_{CQL}	1.0	
Discount factor γ	0.99	
Patient Similarity	Number of clusters K	10
Kernel weights (w_d, w_c, w_m)	(0.2, 0.5, 0.3)	
Transfer blending λ	0.1	
Network	Architecture	MLP, 2 hidden layers × 64 units, ReLU
Evaluation	Independent runs	30

All experiments are conducted on a workstation equipped with an NVIDIA A100 GPU (40 GB), AMD EPYC 7763 processor (64 cores), and 256 GB RAM. The framework is implemented in Python 3.10 using PyTorch 2.1 for neural network components and a custom DE implementation optimized with NumPy vectorized operations.

5 | Results

5.1 | Type 2 Diabetes Management Results

Table 2 presents the comprehensive results for T2D management across all compared methods. METRO-RL achieves the highest composite glycemic control score (0.782 ± 0.024), representing a 16.2% relative improvement over standard DQN (0.673 ± 0.031) and an 18.8% relative improvement over the clinician policy (0.658 ± 0.035). The mean HbA1c reduction achieved by METRO-RL (1.82% ± 0.14%) substantially exceeds that of the best standalone RL baseline, CQL (1.56% ± 0.18%), while maintaining a lower hypoglycemia rate (3.2% vs. 4.1%).

Table 2. T2D management results (mean ± standard deviation over 30 runs). Bold indicates the best result; underline indicates the second-best result.

Method	Composite Score ↑	HbA1c Reduction (%) ↑	Hypo. Rate (%) ↓
METRO-RL (ours)	0.782 ± 0.024	1.82 ± 0.14	3.2 ± 0.4
CQL	0.711 ± 0.028	1.56 ± 0.18	4.1 ± 0.6
BCQ	0.694 ± 0.030	1.48 ± 0.20	3.8 ± 0.5
Dueling DQN	0.689 ± 0.033	1.45 ± 0.19	4.5 ± 0.7
Standard DQN	0.673 ± 0.031	1.38 ± 0.21	4.8 ± 0.8
A2C	0.651 ± 0.038	1.30 ± 0.24	5.3 ± 0.9
Clinician Policy	0.658 ± 0.035	1.24 ± 0.16	5.1 ± 0.6
OpenAI ES	0.641 ± 0.042	1.28 ± 0.26	5.6 ± 1.1
CMA-ES	0.636 ± 0.039	1.25 ± 0.23	5.9 ± 1.0
Random Policy	0.312 ± 0.055	0.41 ± 0.32	12.4 ± 2.3

Notably, METRO-RL achieves both superior efficacy (higher HbA1c reduction) and superior safety (lower hypoglycemia rate) simultaneously, a combination that proves elusive for standalone RL methods, which typically improve efficacy at the cost of increased adverse events. The pure evolutionary methods (OpenAI ES, CMA-ES) perform worse than the offline RL methods, suggesting that gradient-based refinement is essential for achieving high-quality policies in this domain, but that evolutionary initialization (as in METRO-RL) provides a crucial advantage over random initialization.

5.2|Sepsis Treatment Results

Table 3 presents the sepsis treatment results on the MIMIC-IV test set. METRO-RL achieves the lowest estimated 90-day mortality (18.7% ± 2.1%), representing an 11.4% relative reduction compared to the clinician policy (24.3% ± 1.9%) and a 7.0% relative reduction compared to CQL (20.1% ± 2.5%). The estimated SOFA score reduction is also most favorable for METRO-RL (−3.8 ± 0.6 points), indicating superior organ function recovery. METRO-RL policies also yield higher estimated vasopressor-free hours (38.2 ± 2.4 out of 48 hours), suggesting that the learned policy achieves better outcomes with less aggressive vasopressor administration.

Table 3. Sepsis treatment results on MIMIC-IV test set (mean ± SD over 30 runs). Mortality and ICU LOS: lower is better. SOFA reduction and VP-free hours: higher magnitude is better.

Method	Estimated 90-Day Mortality (%) ↓	SOFA Reduction ↑	VP-Free Hours ↑	Estimated ICU LOS (Days) ↓
METRO-RL (ours)	18.7 ± 2.1	−3.8 ± 0.6	38.2 ± 2.4	4.1 ± 0.8
CQL	20.1 ± 2.5	−3.4 ± 0.7	35.6 ± 2.8	4.6 ± 0.9
BCQ	20.8 ± 2.6	−3.2 ± 0.7	36.1 ± 2.6	4.4 ± 0.9
Dueling DQN	21.5 ± 2.9	−3.0 ± 0.8	34.8 ± 3.1	4.8 ± 1.0
Standard DQN	22.4 ± 2.8	−2.8 ± 0.9	33.5 ± 3.4	5.1 ± 1.1
A2C	23.1 ± 3.2	−2.5 ± 1.0	32.7 ± 3.5	5.3 ± 1.2
Clinician policy	24.3 ± 1.9	−2.6 ± 0.5	31.4 ± 2.0	5.5 ± 0.8
OpenAI ES	24.8 ± 3.5	−2.3 ± 1.1	30.9 ± 3.8	5.7 ± 1.3
CMA-ES	25.2 ± 3.3	−2.1 ± 1.0	30.1 ± 3.6	5.9 ± 1.4
Random policy	38.6 ± 4.2	−0.4 ± 1.5	22.3 ± 5.1	7.8 ± 2.0

The performance gap between METRO-RL and the clinician policy is particularly notable given that sepsis management is a domain where experienced intensivists already employ sophisticated, context-dependent treatment strategies. The 5.6 percentage point absolute reduction in estimated mortality suggests that METRO-RL identifies treatment patterns that are systematically superior to historical practice, particularly in the timing and dosing of vasopressor initiation and fluid management during the critical early hours of sepsis treatment.

5.3|Off-Policy Evaluation Results

Table 4 presents the OPE results using three complementary estimators. The consistency of METRO-RL's superiority across all three OPE methods provides strong evidence that the observed performance gains are not artifacts of a particular estimation methodology.

Table 4. OPE estimates on the sepsis test set (mean ± SD over 30 runs). Higher values indicate better estimated policy performance.

Method	WIS Value ↑	FQE Value ↑	DR Value ↑
METRO-RL (Ours)	8.42 ± 0.63	8.87 ± 0.48	8.61 ± 0.52
CQL	7.56 ± 0.78	8.01 ± 0.61	7.74 ± 0.67
BCQ	7.21 ± 0.82	7.68 ± 0.65	7.39 ± 0.71
Dueling DQN	6.89 ± 0.91	7.35 ± 0.72	7.08 ± 0.78
Standard DQN	6.52 ± 0.95	7.04 ± 0.78	6.71 ± 0.84
A2C	6.18 ± 1.12	6.72 ± 0.85	6.41 ± 0.93
Clinician Policy	6.35 ± 0.54	6.35 ± 0.54	6.35 ± 0.54
OpenAI ES	5.94 ± 1.24	6.48 ± 0.92	6.15 ± 1.05
CMA-ES	5.71 ± 1.18	6.22 ± 0.88	5.91 ± 1.01

METRO-RL not only achieves the highest estimated value across all OPE methods but also exhibits the lowest variance, with standard deviations 20–40% smaller than those of standalone DQN methods. This reduced variance reflects the stabilizing effect of DE-based initialization: by starting from a globally optimized policy parameter region, the subsequent DQN refinement is less sensitive to random seed and mini-batch sampling effects. The consistency across WIS, FQE, and DR estimates is encouraging, as disagreement among estimators would indicate potential unreliability in the evaluation. The Spearman rank correlation between WIS and FQE rankings across methods is 0.98, and between FQE and DR is 0.97, indicating strong agreement among the estimators.

5.4 | Safety Analysis

Table 5 presents the safety analysis comparing contraindicated action rates and treatment volatility across all methods. METRO-RL achieves the lowest rate of contraindicated actions (0.3% on T2D, 0.4% on sepsis), demonstrating that the integrated safety framework combining action masking, clinical constraint-masked crossover in DE, and the Lagrangian penalty effectively prevents dangerous treatment recommendations. The residual 0.3–0.4% contraindication rate arises from edge cases at the boundary of the masking criteria (e.g., patients transitioning into hypoglycemia within a decision interval).

Table 5. Safety analysis: contraindicated action rate (%) and treatment volatility index (normalized to clinician policy = 1.00).

Method	Contraindicated Actions (%) ↓		Volatility Index ↓	
	T2D	Sepsis	T2D	Sepsis
METRO-RL (Ours)	0.3	0.4	0.82	0.91
CQL	0.8	1.1	0.95	0.98
BCQ	1.2	1.5	0.97	1.02
Standard DQN	2.1	2.8	1.34	1.45
Dueling DQN	1.8	2.4	1.28	1.38
A2C	3.5	4.2	1.52	1.61
OpenAI ES	4.7	5.3	1.78	1.85
CMA-ES	5.1	5.8	1.83	1.91
Clinician Policy	0.5	0.6	1.00	1.00

The treatment volatility index reveals another important advantage of METRO-RL: its policies recommend smoother treatment transitions than both standalone RL methods and evolutionary approaches. The volatility index of 0.82 (T2D) and 0.91 (sepsis) indicates that METRO-RL produces treatment trajectories that are 8–18% smoother than clinician behavior. This reduced volatility is clinically desirable, as frequent, large changes in medication dosing are associated with adverse events, reduced patient adherence, and increased nursing workload. In contrast, standard DQN exhibits a volatility index of 1.34–1.45, and evolutionary methods show even higher volatility (1.78–1.91), reflecting the lack of temporal consistency in policies that are not explicitly optimized for smooth treatment transitions.

5.5 | Ablation Study

Table 6 presents the ablation study results, quantifying the contribution of each METRO-RL component to overall performance. Results are reported for the sepsis domain (MIMIC-IV test set), with similar patterns observed for T2D (omitted for brevity).

Table 6. Ablation study on the sepsis task (MIMIC-IV test set). Each row removes one component from the full METRO-RL framework.

Configuration	Est. 90-day Mortality (%) ↓	FQE Value ↑	Contraindicated (%) ↓
Full METRO-RL	18.7 ± 2.1	8.87 ± 0.48	0.4
− DE Stage (DQN only, random init)	21.8 ± 3.0	7.42 ± 0.71	1.6
− DQN Stage (DE only, no refinement)	22.5 ± 3.4	6.95 ± 0.86	3.2
− Patient Similarity Kernel	20.3 ± 2.5	8.24 ± 0.55	0.5
− Safety Constraints	19.1 ± 2.3	8.74 ± 0.51	4.8
− CQL Penalty	19.8 ± 2.8	8.51 ± 0.59	0.5
− State-Aware Mutation	19.9 ± 2.6	8.45 ± 0.56	0.5

The ablation reveals several important findings. First, both stages are essential: removing the DE stage (using random initialization for DQN) increases estimated mortality by 3.1 percentage points, while removing the DQN stage (using DE-only policies) increases mortality by 3.8 percentage points. The ablation results confirm the complementary nature of global search and local refinement. Second, the patient similarity kernel contributes a 1.6 percentage point mortality reduction, validating the benefit of knowledge transfer across similar patient subgroups. Third, removing safety constraints slightly improves the estimated mortality (by 0.4 percentage points), reflecting the inherent trade-off between safety and aggressive optimization; however, the contraindicated action rate increases dramatically from 0.4% to 4.8%, making this configuration clinically unacceptable. Fourth, the CQL penalty and state-aware mutation each contribute approximately 1 percentage point of mortality reduction, demonstrating their individual value within the framework.

5.6 | Patient Subgroup Analysis

To assess whether METRO-RL's advantages are uniform across patient populations or concentrated in specific subgroups, we stratify the sepsis test set by baseline disease severity into three groups based on admission SOFA score: mild (SOFA 2–5, n = 624), moderate (SOFA 6–9, n = 781), and severe (SOFA ≥10, n = 445).

Table 7. Estimated 90-day mortality (%) by patient severity subgroup for the top four methods.

Method	Mild (SOFA 2–5)	Moderate (SOFA 6–9)	Severe (SOFA ≥10)
METRO-RL	7.2 ± 1.8	16.5 ± 2.4	34.8 ± 3.6
CQL	8.1 ± 2.1	18.4 ± 2.8	37.6 ± 4.2
BCQ	8.5 ± 2.3	19.2 ± 3.0	38.9 ± 4.5
Clinician	9.8 ± 1.5	22.1 ± 2.2	44.7 ± 3.1

METRO-RL's advantage is most pronounced in the severe subgroup, where it achieves an estimated 34.8% mortality compared to 44.7% for the clinician policy, a 9.9 percentage point absolute reduction. This finding is clinically significant because severely ill patients represent the population with the greatest unmet therapeutic need and the highest potential for benefit from optimized treatment policies. The patient-state-aware mutation mechanism in *Stage 1*, which increases exploration intensity for severe states, likely contributes to this disproportionate benefit by encouraging the discovery of unconventional but effective treatment strategies for the most critically ill patients. For mild patients, the advantage is more modest (2.6 percentage point reduction), consistent with the expectation that standard clinician practice is already reasonably effective for lower-acuity patients.

5.7 | Computational Cost

Table 8 reports the wall-clock training time for each method on the sepsis dataset, measured on the hardware described in Section 4.5.

Table 8. Wall-clock training time comparison (single run, sepsis dataset).

Method	Training Time
METRO-RL (Stage 1 + Stage 2)	2 h 45 minute (2 h DE + 45 minute DQN)
Standard DQN	32 minute
Dueling DQN	35 minute
BCQ	48 minute
CQL	42 minute
A2C	38 minute
OpenAI ES	3 h 10 minute
CMA-ES	4 h 05 minute

METRO-RL's total training time of approximately 2 hours and 45 minutes is roughly 5 times longer than standalone DQN methods but substantially faster than CMA-ES (4 hours 5 minutes) and comparable to OpenAI ES (3 hours 10 minutes). The computational overhead is concentrated in *Stage 1* (DE), where 100 candidate policies must be evaluated across 200 generations. Importantly, the DE stage is embarrassingly

parallel: each candidate policy evaluation is independent, and the current implementation uses 16 parallel workers for fitness evaluation, reducing the effective time per generation. With full utilization of the 64-core processor, *Stage 1* could be further accelerated to under 1 hour. Given that METRO-RL training is a one-time offline computation, the 2.75-hour training time is negligible compared to the clinical timescales of the target applications (months for T2D, days for sepsis).

6 | Discussion

The experimental results demonstrate that METRO-RL's hybrid metaheuristic–RL architecture consistently outperforms both pure RL and pure evolutionary approaches for clinical policy optimization across two distinct medical domains. In this section, we discuss the mechanistic reasons for this superiority, address the framework's limitations, and outline directions for future research.

6.1 | Why the Hybrid Approach Outperforms Standalone Methods

The fundamental advantage of METRO-RL lies in the complementary nature of its two stages. DE's population-based search provides effective global exploration of the policy parameter space, maintaining diversity and avoiding premature convergence to local optima, a well-known failure mode of gradient-based RL methods [55]. The ablation study (*Table 5*) confirms that removing DE initialization and relying on random initialization for DQN increases estimated sepsis mortality by 3.1 percentage points, demonstrating that the gradient-based optimizer alone frequently converges to inferior local optima. Conversely, DE alone (without DQN refinement) produces policies that are coarser and less precise than those refined by gradient-based optimization, as evidenced by the 3.8 percentage point mortality increase when the DQN stage is removed. The DE stage efficiently identifies the "basin of attraction" containing the global optimum, while the DQN stage performs the fine-grained descent within that basin.

6.2 | Role of the Patient Similarity Kernel

The patient similarity kernel serves as a structured prior that regularizes both the DE search and the DQN training by encoding the clinical intuition that similar patients should receive similar treatments. Without this kernel, the framework must independently optimize policies for each patient cluster, wasting computational resources on redundant exploration. The kernel's DTW-based clinical trajectory component is particularly valuable because it captures temporal patterns in disease progression that static demographic features cannot, enabling the identification of patient subgroups that follow similar clinical courses despite potentially different baseline characteristics. The 1.6 percentage point mortality improvement attributable to the kernel (*Table 5*) validates this design choice.

6.3 | Safety and Clinical Deployment Considerations

A critical prerequisite for any AI-based treatment recommendation system is the assurance of patient safety. METRO-RL's multi-layered safety framework combining hard constraint action masking with soft constraint Lagrangian penalties achieves the lowest rate of contraindicated actions among all evaluated methods (0.3–0.4%), approaching the level of human clinician practice (0.5–0.6%). The low residual rate of the clinician policy reflects the reality that even experienced physicians occasionally deviate from guidelines in edge cases. We emphasize that the safety framework is a necessary but not sufficient condition for clinical deployment. Before any real-world use, METRO-RL policies would require: 1) extensive prospective validation in randomized clinical trials, 2) regulatory approval under applicable frameworks (e.g., FDA's Software as a Medical Device guidance), 3) integration into clinical decision support systems that present recommendations to clinicians for review rather than autonomous execution, and 4) continuous monitoring for distributional shift between the training data and the deployment population.

6.4 | Limitations

This study has several important limitations that should be considered when interpreting the results. First, all evaluations rely on offline policy estimation using retrospective data and simulated environments. Despite using a battery of four OPE methods, the true clinical impact of METRO-RL policies can only be determined through prospective clinical trials. OPE methods are known to exhibit bias under violations of the positivity assumption (when the behavior policy assigns zero probability to actions selected by the evaluation policy) and under model misspecification [69]. Second, the T2D dataset is synthetically generated from a simplified disease progression model. While the model is calibrated to UKPDS parameters, it cannot capture the full complexity of real-world T2D management, including patient-specific pharmacokinetics, lifestyle variability, and comorbidity interactions. Third, the discrete action space (20 actions for T2D, 25 for sepsis) is a simplification of the continuous dosing decisions that clinicians make in practice. Finer discretization or continuous action spaces could improve clinical granularity but would also increase the computational and statistical complexity of the optimization. Fourth, the framework does not currently model medication adherence, patient preferences, treatment costs, or social determinants of health, all of which substantially influence real-world treatment outcomes. Fifth, the MIMIC-IV sepsis cohort represents a single tertiary academic medical center, and the generalizability of learned policies to other healthcare systems, patient populations, and clinical practices remains to be established.

6.5 | Ethical Considerations

AI-driven treatment policy optimization raises important ethical concerns. First, there is a risk that policies optimized on historical data may perpetuate or amplify existing disparities in care if the training data reflects systemic biases in treatment allocation across demographic groups [70]. Fairness-aware policy optimization, incorporating group-level equity constraints, is an important direction for future work. Second, the "black box" nature of neural network policies limits clinician understanding of the rationale behind specific treatment recommendations. While our framework does not directly address interpretability, the patient similarity kernel provides partial transparency by linking recommendations to the treatment histories of similar patients. Third, deploying AI-based treatment recommendations requires careful attention to liability, informed consent, and the appropriate division of responsibility between the AI system and the supervising clinician [71].

6.6 | Comparison with Recent Offline RL Advances

Since the inception of this work, several offline RL methods have been proposed that could potentially be integrated into the METRO-RL framework. Decision transformer (Chen et al. [72]) recasts RL as a sequence modeling problem using transformer architectures, while implicit Q-learning (IQL) (Kostrikov et al. [73]) avoids querying out-of-distribution actions entirely by learning Q-values through expectile regression. Diffusion-based offline RL methods (Wang et al. [74]) model complex policy distributions using denoising diffusion models. These methods could replace or complement the DQN component in *Stage 2*, and exploring such integrations represents a promising direction for future work.

6.7 | Future Directions

Several extensions of METRO-RL merit investigation. First, extending to continuous action spaces using actor-critic architectures in *Stage 2* would enable finer-grained dosing recommendations. Second, incorporating real-time adaptation mechanisms would allow policies to update as new patient data becomes available during the treatment course, moving toward fully online personalization. Third, extending the framework to multi-morbidity settings, where patients simultaneously manage multiple chronic conditions with potentially interacting treatments, would enhance clinical applicability.

Fourth, integration with clinical workflow systems, including natural language explanations of treatment recommendations and visualization of expected outcomes under alternative treatment strategies, would improve clinician trust and adoption. Fifth, prospective clinical trials, initially as non-interventional studies

comparing METRO-RL recommendations to actual treatment decisions, would provide the essential evidence base for clinical translation.

7 | Conclusion

This paper introduced METRO-RL, a novel hybrid framework that combines DE global search with DQN local refinement for personalized DTR optimization. The framework addresses fundamental limitations of standalone RL methods for clinical policy optimization, including susceptibility to local optima, sample inefficiency, safety constraint enforcement, and poor handling of patient heterogeneity.

The key innovations of METRO-RL include: 1) a two-stage architecture that leverages DE's global exploration to provide strong initialization for subsequent DQN refinement, 2) a patient-state-aware mutation operator that adapts search intensity to clinical severity, 3) a composite patient similarity kernel integrating demographic, temporal, and comorbidity information to enable knowledge transfer across patient subgroups, and 4) a comprehensive safety framework combining action masking and Lagrangian relaxation within a constrained MDP formulation.

Extensive evaluation on T2D management (simulated cohort) and sepsis treatment (MIMIC-IV database) demonstrated that METRO-RL consistently outperforms eight baseline methods across all OPE metrics and clinical outcome measures. On T2D, METRO-RL achieved a 16.2% improvement in composite glycemic control score over standard DQN and 18.8% over the clinician policy, while maintaining the lowest hypoglycemia rate. On sepsis, METRO-RL reduced estimated 90-day mortality to 18.7%, an 11.4% relative reduction compared to clinician practice. The framework's advantages were most pronounced for severely ill patients, the subgroup with the greatest clinical need.

While these results are encouraging, we emphasize that METRO-RL's clinical impact remains to be validated through prospective clinical trials. The framework currently relies on offline evaluation and simulated environments, and the gap between estimated and actual policy performance is a critical open question for all RL-based clinical decision support systems. Nevertheless, METRO-RL represents a principled and effective approach to hybridizing metaheuristic and RL methods for healthcare, and we believe it opens a promising research direction at the intersection of evolutionary computation, deep learning, and personalized medicine.

Ethical Statement

This study was conducted using the publicly available MIMIC-IV database under the PhysioNet Credentialed Data Use Agreement (version 2.2). The T2D dataset was synthetically generated using a computational disease progression model; no real patient data was collected for this component. No direct patient contact or prospective data collection was performed. The institutional ethics boards of King Saud University (IRB-2024-087) and the Indian Institute of Technology Delhi (IITD/IEC/2024/156) reviewed and approved the study protocol, confirming that the use of de-identified retrospective data and simulated data did not require informed consent.

Acknowledgments

This work was supported by the King Saud University Deanship of Scientific Research, Research Group No. RG-1446-087. A. N. was supported by the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India, under Core Research Grant CRG/2024/001234. The authors thank the PhysioNet team for maintaining the MIMIC-IV database and the anonymous reviewers for their constructive feedback.

Author Contributions

R. Al-Q.: conceptualization, methodology, software, writing original draft, supervision, project administration, funding acquisition. A. N.: formal analysis, validation, data curation, writing review & editing, visualization.

Funding

Not applicable.

Data Availability Statement

The MIMIC-IV database is available through PhysioNet (https://physionet.org/content/mimiciv/) upon completion of required training and credentialing. The T2D simulator code, METRO-RL implementation, and experiment scripts are publicly available at https://github.com/qahtani-lab/METRO-RL.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *The new england journal of medicine*, *372*(9), 793. https://doi.org/10.1056/NEJMp1500523

[2] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New england journal of medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[3] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the royal statistical society series b: Statistical methodology*, *65*(2), 331–355. https://doi.org/10.1111/1467-9868.00389

[4] Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle symposium in biostatistics: Analysis of correlated data* (pp. 189-326). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-9076-1_11

[5] Chakraborty, B., & Moodie, E. E. (2013). *Statistical methods for dynamic treatment regimes: Reinforcement learning, causal inference, and personalized medicine*. New York: Springer. https://doi.org/10.1007/978-1-4614-7428-9

[6] Murphy, S. A. (2005). *A generalization error for Q-learning*. https://jmlr.org/papers/volume6/murphy05a/murphy05a.pdf

[7] Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3), 279–292. https://doi.org/10.1007/BF00992698

[8] Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

[9] Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the american statistical association*, *107*(499), 1106–1118. https://doi.org/10.1080/01621459.2012.695674

[10] Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., & Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment: A deep reinforcement learning approach. *Machine learning for healthcare conference* (pp. 147-163). PMLR. https://doi.org/10.48550/arXiv.1705.08422

[11] Padmanabhan, R., Meskin, N., & Haddad, W. M. (2017). Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Mathematical biosciences*, *293*, 11–20. https://doi.org/10.1016/j.mbs.2017.08.004

[12] Shortreed, S. M., Laber, E., Lizotte, D. J., Stroup, T. S., Pineau, J., & Murphy, S. A. (2011). Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine learning*, *84*(1), 109–136. https://doi.org/10.1007/s10994-010-5229-0

[13] Daskalaki, E., Diem, P., & Mougiakakou, S. G. (2013). An actor-critic based controller for glucose regulation in type 1 diabetes. *Computer methods and programs in biomedicine*, *109*(2), 116–125. https://doi.org/10.1016/j.cmpb.2012.03.002

[14] Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine*, *25*(1), 16–18. https://doi.org/10.1038/s41591-018-0310-5

[15] Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). *Offline reinforcement learning: Tutorial, review, and perspectives on open problems*. https://doi.org/10.48550/arXiv.2005.01643

[16] Fujimoto, S., Meger, D., & Precup, D. (2019). Off-policy deep reinforcement learning without exploration. *International conference on machine learning* (pp. 2052-2062). PMLR. https://proceedings.mlr.press/v97/fujimoto19a/fujimoto19a.pdf

[17] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, *33*, 1179–1191. https://dl.acm.org/doi/abs/10.5555/3495724.3495824

[18] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. *International conference on machine learning* (pp. 22-31). PMLR. https://proceedings.mlr.press/v70/achiam17a.html

[19] Storn, R., & Price, K. (1997). Differential evolution-A simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

[20] Das, S., & Suganthan, P. N. (2010). Differential evolution: A survey of the state-of-the-art. *IEEE transactions on evolutionary computation*, *15*(1), 4–31. https://doi.org/10.1109/TEVC.2010.2059031

[21] Petrovski, A., & McCall, J. (2001). Multi-objective optimisation of cancer chemotherapy using evolutionary algorithms. *International conference on evolutionary multi-criterion optimization* (pp. 531-545). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44719-9_37

[22] Holdsworth, C., Kim, M., Liao, J., & Phillips, M. H. (2010). A hierarchical evolutionary algorithm for multiobjective optimization in IMRT. *Medical physics*, *37*(9), 4986–4997. https://doi.org/10.1118/1.3478276

[23] Abo-Hammour, Z. S., Samhouri, A. D., & Mubarak, Y. (2014). Continuous genetic algorithm as a novel solver for Stokes and nonlinear Navier Stokes problems. *Mathematical problems in engineering*, *2014*(1), 649630. https://doi.org/10.1155/2014/649630

[24] Stanley, K. O., Clune, J., Lehman, J., & Miikkulainen, R. (2019). Designing neural networks through neuroevolution. *Nature machine intelligence*, *1*(1), 24–35. https://doi.org/10.1038/s42256-018-0006-z

[25] Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., ... & Kavukcuoglu, K. (2017). *Population based training of neural networks*. https://doi.org/10.48550/arXiv.1711.09846

[26] Salimans, T., Ho, J., Chen, X., Sidor, S., & Sutskever, I. (2017). *Evolution strategies as a scalable alternative to reinforcement learning*. https://doi.org/10.48550/arXiv.1703.03864

[27] Khadka, S., & Tumer, K. (2018). Evolution-guided policy gradient in reinforcement learning. *Advances in neural information processing systems*, *31*. https://dl.acm.org/doi/10.5555/3326943.3327053

[28] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9–12), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6

[29] Zhao, Y., Zeng, D., Socinski, M. A., & Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, *67*(4), 1422–1433. https://doi.org/10.1111/j.1541-0420.2011.01572.x

[30] Moodie, E. E. M., Dean, N., & Sun, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in biosciences*, *6*(2), 223–243. https://doi.org/10.1007/s12561-013-9103-z

[31] Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A" SMART" design for building individualized treatment sequences. *Annual review of clinical psychology*, *8*(1), 21–48. https://doi.org/10.1146/annurev-clinpsy-032511-143152

[32] Almirall, D., Nahum-Shani, I., Sherwood, N. E., & Murphy, S. A. (2014). Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Translational behavioral medicine*, *4*(3), 260–274. https://doi.org/10.1007/s13142-014-0265-0

[33] Zhang, B., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, *100*(3), 681–694. https://doi.org/10.1093/biomet/ast014

[34] Zhao, Y.-Q., Zeng, D., Laber, E. B., & Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the american statistical association*, *110*(510), 583–598. https://doi.org/10.1080/01621459.2014.937488

[35] Laber, E. B., & Zhao, Y.Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, *102*(3), 501–514. https://doi.org/10.1093/biomet/asv028

[36] Zhang, Y., Laber, E. B., Davidian, M., & Tsiatis, A. A. (2018). Interpretable dynamic treatment regimes. *Journal of the american statistical association*, *113*(524), 1541–1549. https://doi.org/10.1080/01621459.2017.1345743

[37] Tsiatis, A. A., Davidian, M., Holloway, S. T., & Laber, E. B. (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. CRC Press. https://doi.org/10.1201/9780429192692

[38] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, *24*(11), 1716–1720. https://doi.org/10.1038/s41591-018-0213-5

[39] Prasad, N., Cheng, L.F., Chivers, C., Draugelis, M., & Engelhardt, B. E. (2017). *A reinforcement learning approach to weaning of mechanical ventilation in intensive care units*. https://doi.org/10.48550/arXiv.1704.06300

[40] Nemati, S., Ghassemi, M. M., & Clifford, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2978-2981). IEEE. https://doi.org/10.1109/EMBC.2016.7591355

[41] Fox, I., Lee, J., Pop-Busui, R., & Wiens, J. (2020). Deep reinforcement learning for closed-loop blood glucose control. *Machine learning for healthcare conference* (pp. 508-536). PMLR. https://proceedings.mlr.press/v126/fox20a.html

[42] Wu, Y., Tucker, G., & Nachum, O. (2019). *Behavior regularized offline reinforcement learning*. https://doi.org/10.48550/arXiv.1911.11361

[43] Tessler, C., Mankowitz, D. J., & Mannor, S. (2018). *Reward constrained policy optimization*. https://doi.org/10.48550/arXiv.1805.11074

[44] Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). *Adversarial attacks on neural network policies*. https://doi.org/10.48550/arXiv.1702.02284

[45] Hanset, A., Meskens, N., & Duvivier, D. (2010). Using constraint programming to schedule an operating theatre. *2010 IEEE workshop on health care management (WHCM)* (pp. 1-6). IEEE. https://doi.org/10.1109/WHCM.2010.5441245

[46] Anter, A. M., & Hassenian, A. E. (2019). CT liver tumor segmentation hybrid approach using neutrosophic sets, fast fuzzy c-means and adaptive watershed algorithm. *Artificial intelligence in medicine*, *97*, 105–117. https://doi.org/10.1016/j.artmed.2018.11.007

[47] Li, Y., Yao, J., & Yao, D. (2004). Automatic beam angle selection in IMRT planning using genetic algorithm. *Physics in medicine & biology*, *49*(10), 1915–1932. https://doi.org/10.1088/0031-9155/49/10/007

[48] Veng-Pedersen, P., Gobburu, J. V. S., Meyer, M. C., & Straughn, A. B. (2000). Carbamazepine level-A in vivo-in vitro correlation (IVIVC): A scaled convolution based predictive approach. *Biopharmaceutics & drug disposition*, *21*(1), 1–6. https://doi.org/10.1002/1099-081X(200001)21:1%3C1::AID-BDD207%3E3.0.CO;2-D

[49] Marchetti, G., Barolo, M., Jovanovic, L., Zisser, H., & Seborg, D. E. (2008). An improved PID switching control strategy for type 1 diabetes. *IEEE transactions on biomedical engineering*, *55*(3), 857–865. https://doi.org/10.1109/TBME.2008.915665

[50] Moles, C. G., Mendes, P., & Banga, J. R. (2003). Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome research*, *13*(11), 2467–2474. https://doi.org/10.1101/gr.1262503

[51] Riquelme, N., Von Lücken, C., & Baran, B. (2015). Performance metrics in multi-objective optimization. *2015 Latin American computing conference (CLEI)* (pp. 1-11). IEEE. https://doi.org/10.1109/CLEI.2015.7360024

[52] Stanley, K. O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary computation*, *10*(2), 99–127. https://doi.org/10.1162/106365602320169811

[53] Stanley, K. O., D'Ambrosio, D. B., & Gauci, J. (2009). A hypercube-based encoding for evolving large-scale neural networks. *Artificial life*, *15*(2), 185–212. https://doi.org/10.1162/artl.2009.15.2.15202

[54] Pourchot, A., & Sigaud, O. (2018). *CEM-RL: Combining evolutionary and gradient-based methods for policy search*. https://doi.org/10.48550/arXiv.1810.01222

[55] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *Proceedings of the AAAI conference on artificial intelligence* (pp. 3207–3214). AAAI Press. https://doi.org/10.1609/aaai.v32i1.11694

[56] Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., & Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. *International conference on machine learning* (pp. 1995-2003). PMLR. https://proceedings.mlr.press/v48/wangf16.html

[57] Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). *Prioritized experience replay*. https://doi.org/10.48550/arXiv.1511.05952

[58] Precup, D., Sutton, R. S., & Singh, S. (2000). Eligibility traces for off-policy policy evaluation. *Proceedings of the seventeenth international conference on machine learning (ICML)* (pp. 759–766). Morgan Kaufmann. https://hdl.handle.net/20.500.14394/10401

[59] Le, H., Voloshin, C., & Yue, Y. (2019). Batch policy learning under constraints. *Proceedings of the 36th international conference on machine learning*, *97*, 3703–3712. https://proceedings.mlr.press/v97/le19a.html

[60] Jiang, N., & Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. *Proceedings of the 33rd international conference on machine learning* (PP. 652–661). PMLR. https://proceedings.mlr.press/v48/jiang16.pdf

[61] Kovatchev, B. P., Breton, M., Dalla Man, C., & Cobelli, C. (2009). In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes. *Journal of Diabetes science and technology*, *3*(1), 44–55. https://doi.org/10.1177/193229680900300106

[62] Man, C. D., Micheletto, F., Lv, D., Breton, M., Kovatchev, B., & Cobelli, C. (2014). The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes science and technology*, *8*(1), 26–34. https://doi.org/10.1177/1932296813514502

[63] Holman, R. R., Paul, S. K., Bethel, M. A., Matthews, D. R., & Neil, H. A. W. (2008). 10-year follow-up of intensive glucose control in type 2 diabetes. *New england journal of medicine*, *359*(15), 1577–1589. https://doi.org/10.1056/NEJMoa0806470

[64] Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., ... & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, *10*(1), 1. https://doi.org/10.1038/s41597-022-01899-x

[65] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., ... & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, *315*(8), 801-810. https://doi.org/10.1001/jama.2016.0287

[66] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). *Playing Atari with deep reinforcement learning*. https://doi.org/10.48550/arXiv.1312.5602

[67] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *Proceedings of the 33rd international conference on machine learning* (pp. 1928–1937). PMLR. https://proceedings.mlr.press/v48/mniha16.html

[68] Hansen, N. (2006). The CMA evolution strategy: A comparing review. *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms*, 75–102. https://doi.org/10.1007/3-540-32494-1_4

[69] Thomas, P., & Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. *International conference on machine learning* (pp. 2139-2148). PMLR. https://proceedings.mlr.press/v48/thomasa16.html

[70] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

[71] Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, *46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

[72] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., ... & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, *34*, 15084-15097. https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html

[73] Kostrikov, I., Nair, A., & Levine, S. (2021). *Offline reinforcement learning with implicit q-learning*. https://doi.org/10.48550/arXiv.2110.06169

[74] Wang, Z., Hunt, J. J., & Zhou, M. (2022). *Diffusion policies as an expressive policy class for offline reinforcement learning*. https://doi.org/10.48550/arXiv.2208.06193