ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

A Risk-Aware Evaluation Framework for Reinforcement Learning-Based Adaptive Cancer Therapy

Mohammed Umar Alhaji[1]; Usman Ahmad Ahmad[2]; Nasir Muazu Abba[3]

[1]Department of Computer Science, Usmanu Danfodiyo University, Sokoto, Nigeria.
[2]Herriot Watt University, Edinburgh, Scotland, UK.
[3]Department of Computer Science, Usmanu Danfodiyo University, Sokoto, Nigeria, Nigeria.

Publication Date: 2026/02/02

Abstract: Reinforcement learning has emerged as a promising approach for adaptive cancer therapy due to its ability to optimize sequential treatment decisions under uncertainty. While studies have demonstrated the potential of reinforcement learning to improve simulated treatment outcomes, most evaluations rely primarily on average performance metrics obtained through direct simulation rollouts. Such evaluation practices provide limited insight into uncertainty, robustness, and worst-case behavior, which are critical considerations in safety-sensitive clinical domains. This study proposes a standardized, risk-aware, and uncertainty-sensitive evaluation framework for reinforcement learning based adaptive cancer therapy using simulated tumor environments. A Deep Q Network policy is evaluated against clinically interpretable baselines using multiple performance perspectives, including mean outcomes, worst-case metrics, and tail risk measures based on Conditional Value at Risk. Robustness is further assessed under parameter perturbations and distribution shifts representing aggressive tumor dynamics. Experimental results demonstrate that adaptive reinforcement learning policies achieve tumor control comparable to maximum dose therapy while maintaining controlled risk exposure and stable performance under uncertainty. The findings emphasize that rigorous, risk-sensitive evaluation is essential for drawing reliable conclusions about reinforcement learning based treatment strategies before any real-world deployment.

Keywords: *Reinforcement Learning; Adaptive Cancer Therapy; Risk Aware Evaluation; Conditional Value at Risk; Simulation-Based Evaluation; Deep Q Network.*

How to Cite Mohammed Umar Alhaji; Usman Ahmad Ahmad; Nasir Muazu Abba (2026) A Risk-Aware Evaluation Framework for Reinforcement Learning-Based Adaptive Cancer Therapy. *International Journal of Innovative Science and Research Technology*, 11(1), 2528-2538. https://doi.org/10.38124/ijisrt/26jan1026

I. INTRODUCTION

Reinforcement learning has gained significant attention as a framework for sequential decision making in healthcare, where treatment decisions must adapt over time in response to evolving patient states. Unlike supervised learning approaches, reinforcement learning explicitly optimizes long-term outcomes under uncertainty, making it well-suited for dynamic treatment regimens such as drug dosing, therapy scheduling, and intervention timing. As a result, reinforcement learning has been explored for clinical decision support in critical care, chronic disease management, and personalized medicine.

Among healthcare domains, oncology has emerged as a compelling application area for reinforcement learning. Cancer treatment is sequential, with therapeutic decisions influencing tumor evolution, resistance development, and patient toxicity over extended time horizons. Recent studies have demonstrated that reinforcement learning can discover adaptive treatment strategies that outperform fixed or heuristic dosing schedules in simulated cancer environments. These findings suggest that reinforcement learning has strong potential to support adaptive cancer therapy, a paradigm that seeks to control tumor burden while delaying resistance rather than pursuing maximum tolerated dosing.

A notable example is the work by Eastman *et al.*, which demonstrated that reinforcement learning-derived chemotherapy schedules can outperform classical optimal control approaches and remain robust to patient-specific parameter variations in simulated tumor growth models. Such studies establish reinforcement learning as a promising methodological tool for adaptive cancer therapy and provide a foundation for further research in this area. However, despite growing methodological sophistication and encouraging simulation results, the translation of reinforcement learning into high-stakes clinical domains such

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

as oncology remains limited. A central barrier repeatedly identified in the literature is the lack of rigorous, standardized, and reliable evaluation methodologies for healthcare reinforcement learning policies. This issue is particularly acute in cancer therapy, where unsafe or overly optimistic policy recommendations could have severe consequences.

Although reinforcement learning has been widely applied to adaptive cancer therapy, current evaluation practices remain insufficiently standardized, risk-aware, and uncertainty sensitive, which limits the reliability and interpretability of reported results. Most existing studies evaluate reinforcement learning policies through direct rollout in simulated environments, reporting aggregate metrics such as average tumor burden, survival time, or cumulative reward. While insightful, such rollout-based evaluation provides only a partial assessment of policy quality, as it does not quantify uncertainty, assess worst-case outcomes, or capture inconsistencies across patient subpopulations.

Prior work in healthcare reinforcement learning has demonstrated that different evaluation approaches, including on-policy simulation, off-policy evaluation, and model-based estimation, can lead to different conclusions about policy effectiveness. Moreover, commonly used off-policy evaluation methods are known to suffer from high variance, bias, and sensitivity to distribution shift, especially in complex and nonlinear clinical environments. Despite these challenges, no widely accepted evaluation framework exists for reinforcement learning in adaptive cancer therapy. Oncology-focused studies often adopt evaluation strategies in an ad hoc manner, without systematically examining policy robustness, tail risk, or safety under uncertainty. As a result, it remains unclear whether reported performance gains reflect genuine therapeutic improvements or artifacts of evaluation methodology.

Motivated by these limitations, this study focuses not on proposing a new reinforcement learning algorithm, but on how reinforcement learning policies for adaptive cancer therapy are evaluated. The aim is to develop and assess a standardized, risk-aware, and uncertainty-sensitive evaluation framework for reinforcement learning algorithms in adaptive cancer therapy using simulated cancer environments as a controlled yet realistic testbed. Specifically, the study examines off-policy evaluation methods to assess estimator behavior, applies on-policy rollout evaluation as a reliable performance baseline, characterizes policy performance beyond average outcomes using worst-case and risk-sensitive metrics such as Conditional Value at Risk, and evaluates robustness under parameter perturbations and clinically relevant distribution shifts.

II. RELATED WORKS

➢ *Background*

Reinforcement learning involves learning a mapping from situations to actions to maximize a scalar reward or reinforcement signal. The learner is not told which action to take, as in most forms of machine learning, but instead must discover which actions yield the highest reward by trying them [1]. Reinforcement learning (RL) has emerged as a powerful paradigm for solving sequential decision-making problems, where actions influence future system dynamics and outcomes. Unlike supervised learning approaches that rely on static input-output mappings, RL explicitly models temporal dependencies and long-term objectives, making it suitable for healthcare applications involving dynamic treatment regimens (DTRs), where treatment decisions must adapt over time in response to patient evolution [7], [14].

➢ *Reinforcement Learning in Healthcare*

Over the past decade, RL has been investigated as a tool for clinical decision support across various healthcare domains, including critical care, chronic disease management, and personalized medicine. Recent surveys have highlighted that RL methods are well-aligned with clinical scenarios that require balancing short-term interventions against long-term patient outcomes, such as medication dosing, treatment sequencing, and intervention timing [17], [3], [5].

Systematic reviews published in the last few years emphasize that RL has demonstrated promise in learning adaptive treatment strategies that outperform fixed or guideline-based policies in simulated or retrospective settings [9], [8]. These studies consistently report that RL can capture patient heterogeneity and temporal dynamics more effectively than traditional rule-based or regression-based approaches. However, despite these advances, the same reviews repeatedly note persistent challenges related to policy evaluation, safety, interpretability, and clinical trust. In particular, most healthcare RL studies rely on retrospective or simulated data and lack standardized procedures for assessing the reliability and robustness of learned policies [3], [9]. This limitation becomes critical in high-risk domains such as oncology.

➢ *Reinforcement Learning in Oncology and Cancer Treatment*

Oncology represents a natural and compelling application area for RL due to the sequential nature of cancer treatment. Chemotherapy, radiotherapy, and targeted therapies are administered over extended periods, with each treatment decision influencing tumor evolution, resistance development, and patient toxicity. Consequently, optimal cancer treatment planning can be naturally framed as a Markov decision process, where states represent patient or tumor conditions, actions correspond to treatment choices, and rewards encode clinical objectives such as tumor suppression and toxicity minimization [12].

Early work applying RL to chemotherapy scheduling demonstrated that even relatively simple algorithms, such as Q-learning, could discover dosing policies that outperform static treatment schedules in simulated tumor models [11]. These studies established proof-of-concept evidence that RL could adaptively balance efficiency and toxicity over time.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

Subsequent research extended these ideas using more sophisticated modeling assumptions and learning techniques. Several studies incorporated biologically motivated tumor growth models, continuous state spaces, and more realistic toxicity dynamics, showing that RL-based approaches could derive patient-specific or robust dosing strategies under parameter uncertainty [16], [13].

These works collectively reinforced the potential of RL as a decision-support tool for adaptive cancer therapy.

Despite methodological progress, the evaluation strategies employed across oncology-focused RL studies remain performance-oriented. Most works assess learned policies by directly implementing them in simulation environments and reporting aggregate metrics, such as average tumor reduction, survival time, or cumulative reward. While such evaluations demonstrate feasibility, they do not provide insight into the reliability, risk profile, or uncertainty associated with policy recommendations, factors that are critical for any potential clinical translation.

➢ *Adaptive Cancer Therapy and Robust RL Policies*

Adaptive therapy has gained attention as an alternative to maximum-tolerated-dose strategies, particularly in the context of treatment resistance. Rather than aggressively eliminating tumor cells, adaptive therapy aims to control tumor burden while delaying or preventing the emergence of resistant populations [6]. This paradigm naturally aligns with RL, which can optimize long-term objectives under uncertainty.

A representative and influential study in this area is Reinforcement learning derived chemotherapeutic schedules for robust patient-specific therapy [1]. In this work, the authors formulate chemotherapy scheduling as an RL problem using a mechanistic tumor growth model and demonstrate that RL-derived policies outperform classical optimal control strategies across a range of simulated patient parameter variations.

In this study, robustness refers specifically to the stability of policy performance under parameter perturbations and distribution shifts in simulated patient dynamics, rather than formal guarantees derived from robust Markov decision process theory.

The study highlights the robustness of RL policies to model perturbations, positioning RL as a promising approach for patient-specific adaptive therapy. The strength of this work lies in its rigorous simulation design, biologically interpretable modeling assumptions, and comparative evaluation against established control methods. As such, it serves as a strong anchor for subsequent research on RL-based adaptive cancer therapy. However, its evaluation methodology, like most of the existing literature, focuses primarily on mean performance metrics obtained via simulator rollouts. The study does not examine alternative evaluation methodologies, quantify uncertainty in policy performance, or assess worst-case or risk-sensitive outcomes. This pattern is consistent across much of the adaptive cancer therapy literature: RL is validated primarily through direct simulation outcomes, implicitly assuming that rollout-based evaluation provides a reliable estimate of policy quality. While reasonable in controlled settings, this assumption becomes problematic when considering offline learning, limited data coverage, or safety-critical decision-making.

➢ *Evaluation Challenges in Healthcare Reinforcement Learning*

Parallel to the growth of RL applications in healthcare, a separate body of literature has emerged that highlights the fundamental challenges in policy evaluation, particularly in offline settings where interaction with the real environment is not possible. Off-policy evaluation (OPE) methods such as importance sampling, weighted importance sampling, and doubly robust estimators are commonly used to estimate the value of learned policies from logged data [10], [4].

Recent studies, however, demonstrate that OPE methods can exhibit high variance, bias, and sensitivity to data distribution shift, especially in complex, high-dimensional healthcare environments [15], [3]. Critical analyses show that different evaluation methods can produce contradictory conclusions about which policy is optimal, raising concerns about over-optimistic or misleading performance claims [2]. Table 1 summarizes Reinforcement Learning Studies in Adaptive Cancer Therapy.

Table 1 Summary of RL Studies

Study	Environment	Reinforcement Learning Method	Evaluation Approach	Risk/ Uncertainty
Shen *et al.* (2017)	Simulated tumor	Q-learning	On-policy rollout	No
Eastman *et al.* (2021)	Mechanistic tumor model	RL + optimal control	Rollout	Limited
Sun *et al.* (2023)	Simulated oncology	Deep RL	Rollout	No
This study	Simulated heterogeneous tumors	DQN	Rollout + CVaR + stress testing	Yes

➢ *Synthesis and Research Gap*

The literature establishes three key points. First, reinforcement learning is widely recognized as a suitable and powerful framework for adaptive treatment planning in healthcare and oncology. Second, simulated cancer therapy environments have enabled meaningful progress in demonstrating the feasibility and potential benefits of RL-based adaptive therapy. Third, and critically, evaluation practices have not kept pace with methodological advances, remaining fragmented, performance-centric, and insufficiently aligned with safety-critical clinical requirements.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

Anchored by influential studies such as [1], which demonstrate the promise of RL for robust adaptive cancer therapy, this work argues that the next necessary step is a systematic examination of how RL policies are evaluated. Specifically, there is a clear need for a standardized evaluation framework that integrates multiple evaluation perspectives, including uncertainty quantification and risk-sensitive metrics, while leveraging the availability of ground-truth outcomes in simulated cancer environments.

Addressing this gap is important not only for improving the reliability and comparability of future research but also for preventing over-optimistic conclusions that could restrict the responsible translation of reinforcement learning into clinical oncology.

III. METHODOLOGY

This section describes the methodology used to evaluate reinforcement learning policies for adaptive cancer therapy. A simulation-based experimental design is adopted to enable controlled and reproducible comparison of multiple policy evaluation approaches. The methodology emphasizes robustness, uncertainty, and risk-sensitive assessment to reflect the safety-critical nature of clinical decision support.

A modular research framework is employed that integrates simulation, learning, evaluation, and reporting. The overall architecture of the framework is illustrated in Fig. 1.

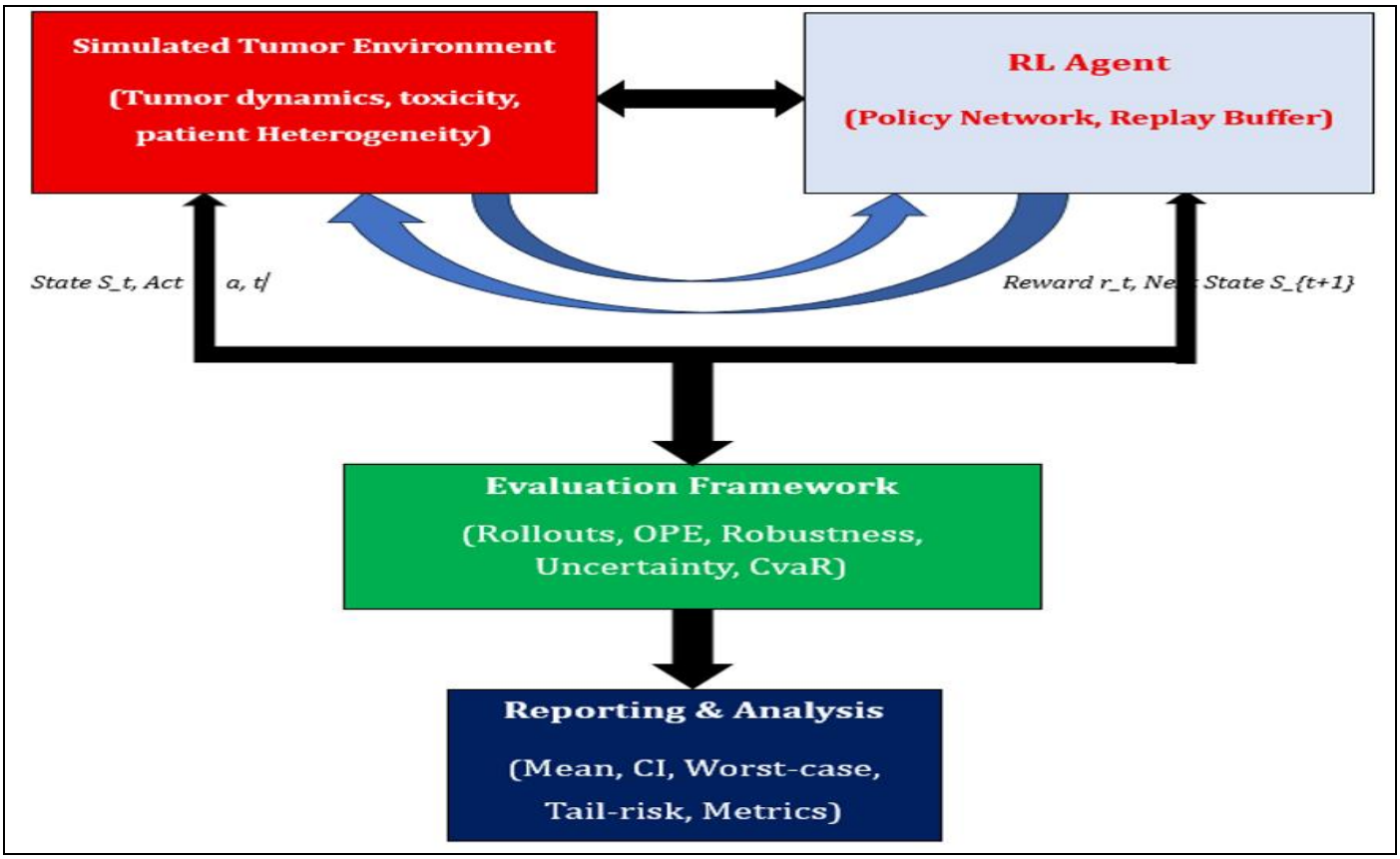

Fig 1 Architectural Diagram

The framework consists of three interacting layers. The learning layer represents the reinforcement learning agent, which selects chemotherapy dosing actions based on observed tumor states and toxicity levels. A Deep Q Network architecture is used, incorporating experience replay and target networks to stabilize learning. The simulation layer models tumor dynamics, resistance evolution, drug effects, and cumulative toxicity. Patient heterogeneity is introduced by sampling biological parameters from clinically plausible distributions. The evaluation layer assesses learned policies using both on-policy rollouts and off-policy evaluation methods, robustness testing under parameter perturbations, and uncertainty quantification using risk-sensitive metrics. Outputs from the evaluation process are summarized into interpretable measures such as mean performance, worst-case outcomes, and tail-risk indicators.

Adaptive cancer therapy is formulated as a discrete-time Markov Decision Process defined by the tuple (S, A, P, R, γ). At each time step, the agent observes the current state, selects a treatment action, receives a reward, and transitions to a new state. The state vector captures clinically relevant tumor and treatment information and is defined as:

$$s_t = [T_s(t), T_r(t), C(t), t] \quad \text{eq.} \quad (1)$$

Where $T_s(t)$ and $T_r(t)$ denote the populations of drug-sensitive and drug-resistant tumor cells, respectively, C(t) represents cumulative drug exposure or toxicity, and t is the

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

treatment time index. This formulation enables explicit modeling of resistance dynamics and aligns with prior adaptive therapy studies.

The action space represents discrete chemotherapy dosing decisions and is defined as:

$A = \{0, d_1, d_2\}$ eq. (2)

Corresponding to no treatment, low-dose treatment, and high-dose treatment. Discrete actions are selected to reflect clinical decision constraints and to maintain comparability with existing literature.

Tumor evolution follows biologically motivated growth dynamics. For a two-population tumor model, the continuous-time dynamics are given by:

$$\frac{dT_s}{dt} = r_s T_s \left(1 - \frac{T_s + T_r}{K}\right) - d_s a_t T_s \quad \text{(eq. 3)}$$

$$\frac{dT_r}{dt} = r_r T_r \left(1 - \frac{T_s + T_r}{K}\right) \quad \text{(eq. 4)}$$

Where r_s and r_r are growth rates, K is the carrying capacity, d_s is drug sensitivity, and a_t denotes the administered dose. Discrete-time transitions are obtained through numerical integration.

The reward function balances tumor suppression and toxicity minimization and is defined as:

$r_t = -\alpha(T_s(t) + T_r(t)) - \beta a_t - \lambda C(t)$ eq. (5)

Where α, β, and λ control the trade-off between tumor suppression, treatment intensity, and cumulative toxicity.

A simulated cancer therapy environment is constructed to emulate patient-specific tumor dynamics. Each episode corresponds to a complete treatment course over a fixed time horizon. Patient heterogeneity is introduced by sampling biological parameters from clinically plausible distributions, with each parameter set representing a distinct virtual patient.

A Deep Q Network is employed as a representative value-based reinforcement learning algorithm. The policy selects actions according to the maximum estimated action-value, and learning is performed using the standard Q-learning update rule with neural network function approximation. Experience replay is used to reduce temporal correlation between samples, and a target network is periodically synchronized to improve training stability.

Training proceeds episodically through interaction between the agent and the simulated environment. At the start of each episode, a new virtual patient profile is sampled. The agent selects actions using an ε-greedy exploration strategy, and observed transitions are stored in a replay buffer. Mini-batches are sampled from the buffer to update the network parameters. Training is conducted across multiple patient profiles and random seeds to promote robustness and generalization. After training, the learned policy is fixed and evaluated using the proposed risk-aware evaluation framework. The training process is illustrated in Fig. 2.

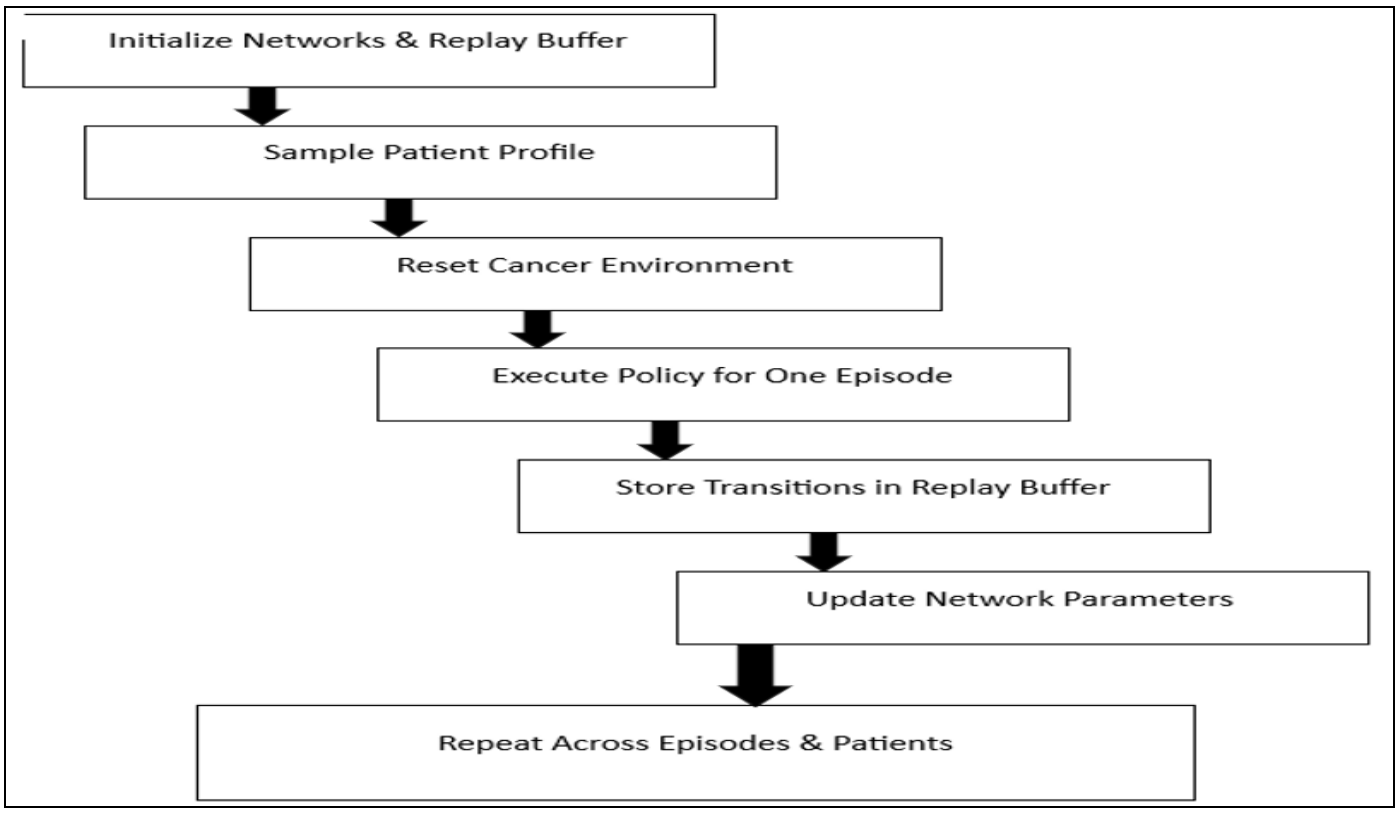

Fig 2 The Training Process

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

IV. RESULTS AND DISCUSSION

This section presents the empirical evaluation of the proposed risk-aware reinforcement learning framework for adaptive cancer therapy. A Deep Q Network (DQN) policy is evaluated against two clinically interpretable baselines: no treatment and maximum tolerated dose (MTD) therapy. Performance is assessed under standard patient conditions and under a distribution shift representing more aggressive tumor dynamics. To ensure clinical relevance, both expected performance and risk-sensitive metrics are reported, including worst-case outcomes and Conditional Value at Risk (CVaR).

All results are obtained from simulations across 50 heterogeneous virtual patients, providing a balance between computational feasibility and distributional coverage consistent with prior simulation-based oncology studies. Although multiple off-policy evaluation methods were implemented, including importance sampling, weighted importance sampling, and doubly robust estimation, their numerical results are not emphasized due to high variance under limited trajectory coverage. Instead, off-policy evaluation methods are used diagnostically to assess estimator stability, while primary conclusions are drawn from controlled on-policy rollouts where ground-truth outcomes are available.

Under standard patient conditions, the DQN policy is compared against no treatment and MTD therapy using mean, median, worst-case, and $CVaR_{0.1}$ final tumor burden metrics, as summarized in Table 2.

Table 2 Final Tumor Burden (Standard)

Policy	Final Tumor Burden (Standard Evaluation)			
	Mean	***Median***	***Worst-case***	***$CvaR_{0,1}$***
No Treatment	1.22×10^{8}	2.33×10	7.38×10^{8}	5.82×10^{8}
Max Dose (MTD)	2.31×10^{6}	6.99×10^{5}	1.79×10^{7}	1.31×10^{7}
DQN	3.00×10^{6}	7.49×10^{5}	1.89×10^{7}	1.46×10^{7}

Relative to no treatment, both MTD and DQN achieve substantial tumor suppression. Mean tumor burden is reduced by approximately 98.1% under MTD and 97.5% under DQN. Worst-case tumor burden is similarly reduced by 97.6% and 97.4%, respectively. These results confirm that active treatment is essential and that both policies effectively control tumor growth.

When comparing DQN to MTD, the learned policy exhibits a 29.8% higher mean tumor burden, while the median tumor burden is only 7.2% higher. Worst-case outcomes increase by approximately 5.4%, and $CVaR_{0.1}$ increases by 11.3%. Despite these increases, all metrics remain within the same order of magnitude, indicating that the DQN policy performs comparably to aggressive fixed-dose therapy. This reflects a fundamental trade-off: the DQN adapts dosing over time rather than consistently applying maximum intensity, which reduces cumulative exposure and potential toxicity.

Distributional analysis further highlights these differences. The no-treatment policy exhibits extreme right skew and heavy tails, corresponding to catastrophic outcomes in a subset of patients. In contrast, both MTD and DQN produce tightly concentrated outcome distributions. The DQN distribution shows slightly greater variance, consistent with adaptive decision-making. This behavior is illustrated in Fig. 3.

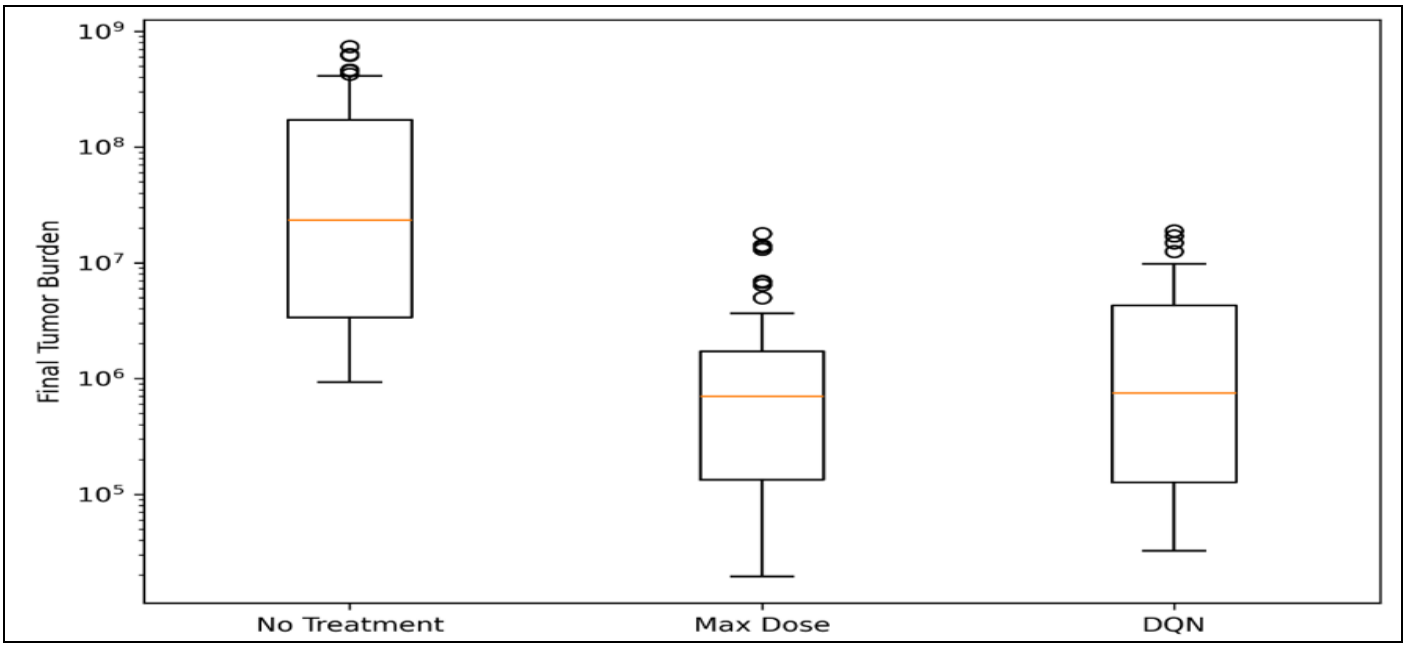

Fig 3 Distribution of Final Tumor Burden Under Standard Conditions

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

Tail-risk analysis using CVaR confirms that no treatment carries catastrophic risk, while both MTD and DQN substantially mitigate extreme outcomes. Although DQN incurs a modest increase in $CVaR_{0.1}$ relative to MTD, it avoids the severe tail behavior observed under no treatment, as shown in Fig. 4.

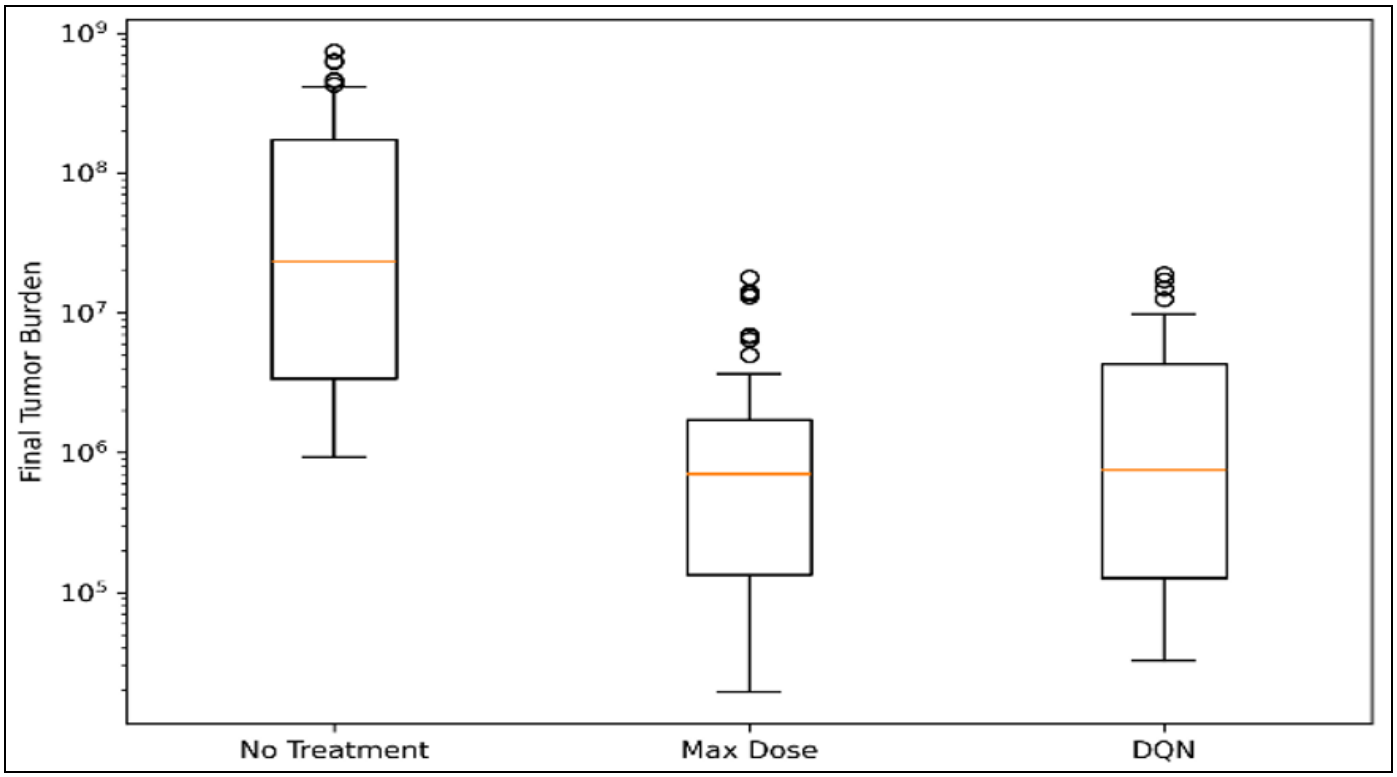

Fig 4 $CVaR_{0.1}$ Comparison Under Standard Conditions.

The relationship between expected performance and worst-case outcomes further illustrates this trade-off. The DQN policy lies close to the Pareto frontier defined by MTD, demonstrating that near-optimal worst-case performance can be achieved without uniform maximal dosing. This is visualized in Fig. 5.

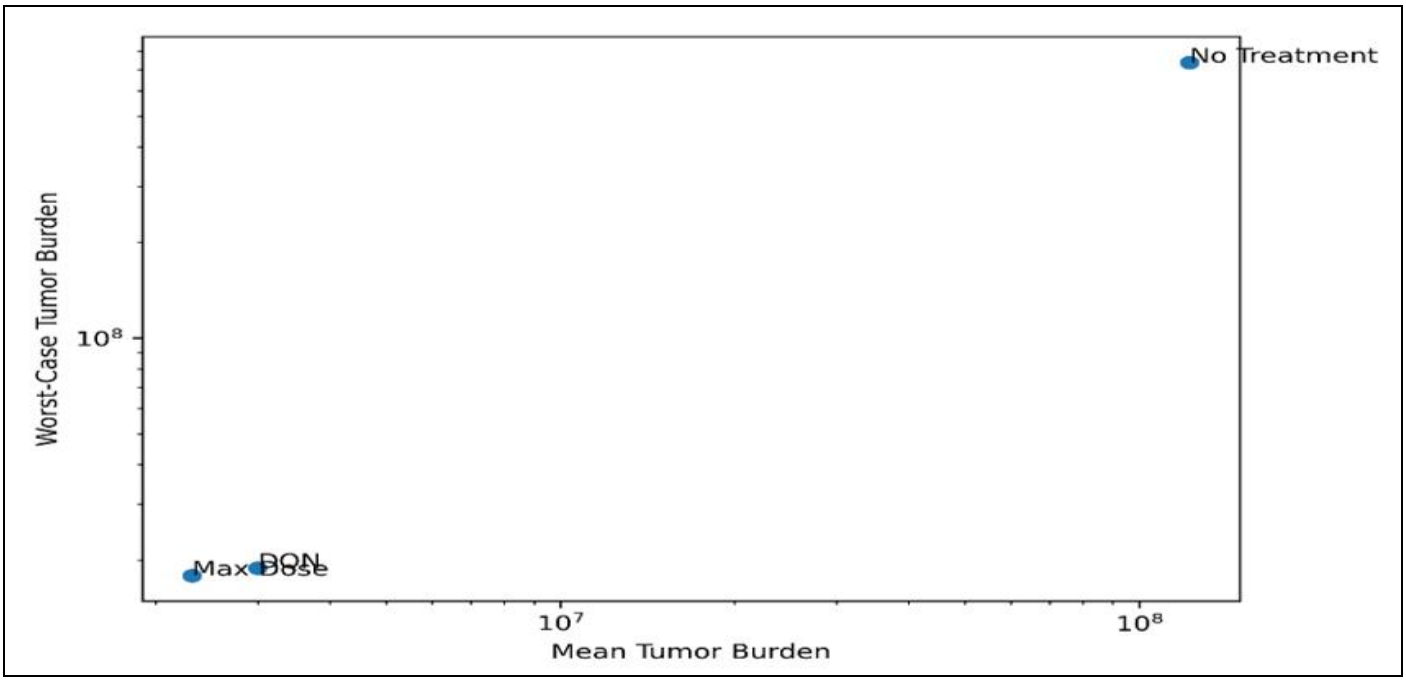

Fig 5 Mean Versus Worst-Case Final Tumor Burden (Standard Conditions)

To assess robustness, all policies are evaluated under a distribution shift representing more aggressive tumor dynamics. Results are summarized in Table 3.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

Table 3 Final Tumor Burden (Aggressive)

Policy	Final Tumor Burden (Aggressive Tumors)			
	Mean	***Median***	***Subhead***	***Subhead***
No Treatment	2.45×10^{8}	4.66×10^{7}	1.48×10^{9}	1.16×10^{9}
Max Dose (MTD)	5.36×10^{6}	9.36×10^{5}	3.26×10^{7}	2.60×10^{7}
DQN	5.76×10^{6}	9.87×10^{5}	3.54×10^{7}	2.78×10^{7}

Under this shift, mean tumor burden increases by 132% for MTD and 92% for DQN, while worst-case outcomes increase by 82% and 88%, respectively. $CVaR_{0.1}$ approximately doubles for both active treatment strategies. Despite this degradation, both MTD and DQN maintain tumor burdens that are two orders of magnitude lower than no treatment.

Relative to MTD under aggressive tumor dynamics, the DQN policy exhibits a 7.4% higher mean tumor burden, a 5.5% higher median, an 8.6% higher worst-case outcome, and a 7.0% higher $CVaR_{0.1}$. These small margins indicate that the learned policy generalizes well beyond its training distribution. Distributional analysis shows a modest widening of the DQN outcome distribution under shift, but without catastrophic tail expansion, demonstrating robustness to parameter uncertainty (Fig. 6).

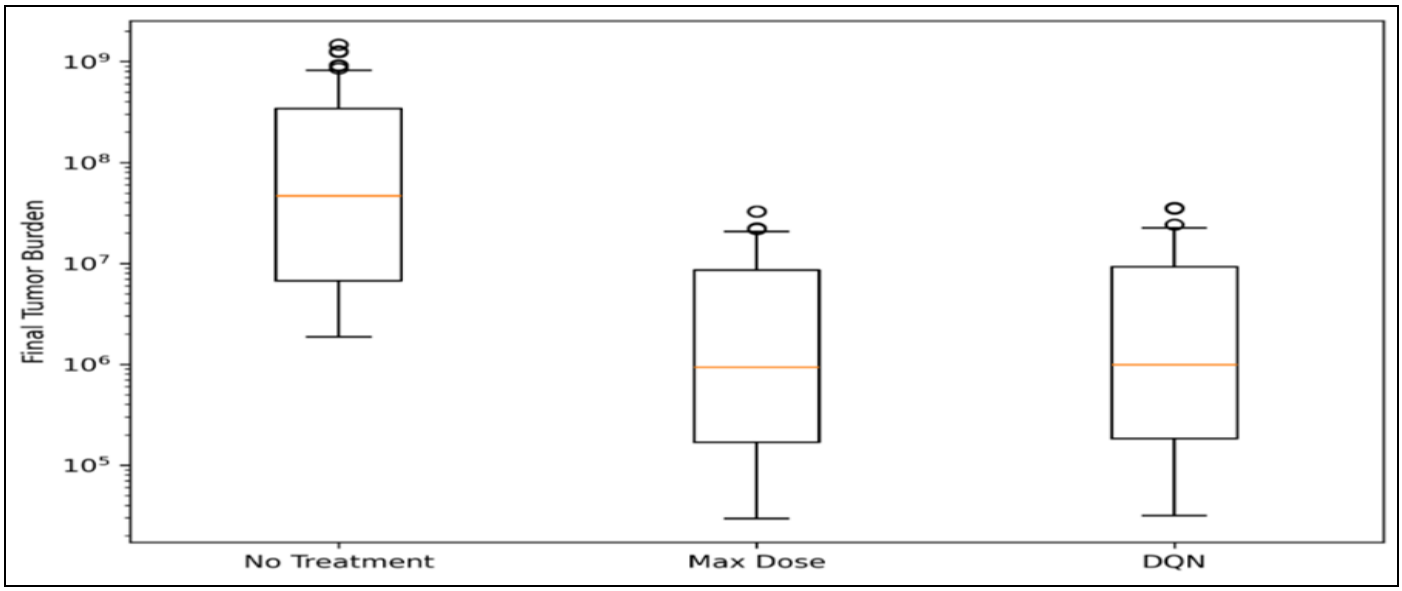

Fig 6 Distribution of Final Tumor Burden Under Aggressive Tumor Dynamics

Risk-sensitive analysis confirms that while tail risk increases for all policies, the DQN maintains controlled risk exposure under stress (Fig. 7).

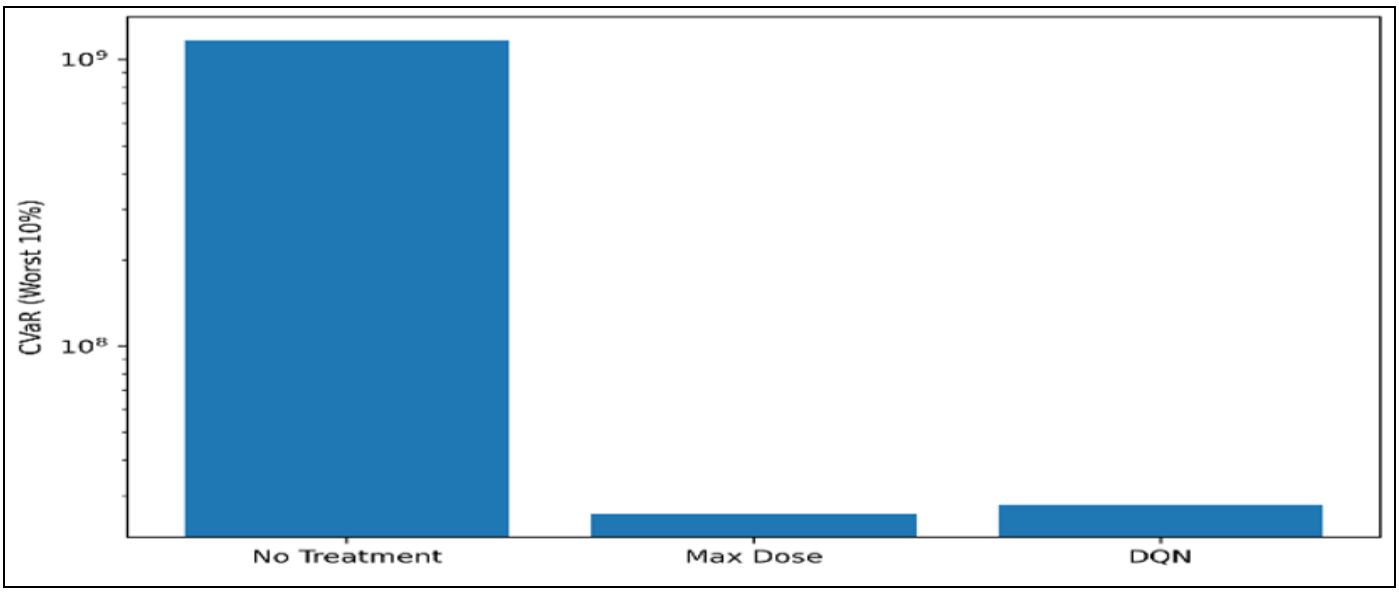

Fig 7 $CVaR_{0,\,1}$ Comparison Under Aggressive Tumors

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

The mean versus worst-case performance under distribution shift further shows that the learned policy continues to occupy a favorable region of the performance-risk space (Fig. 8).

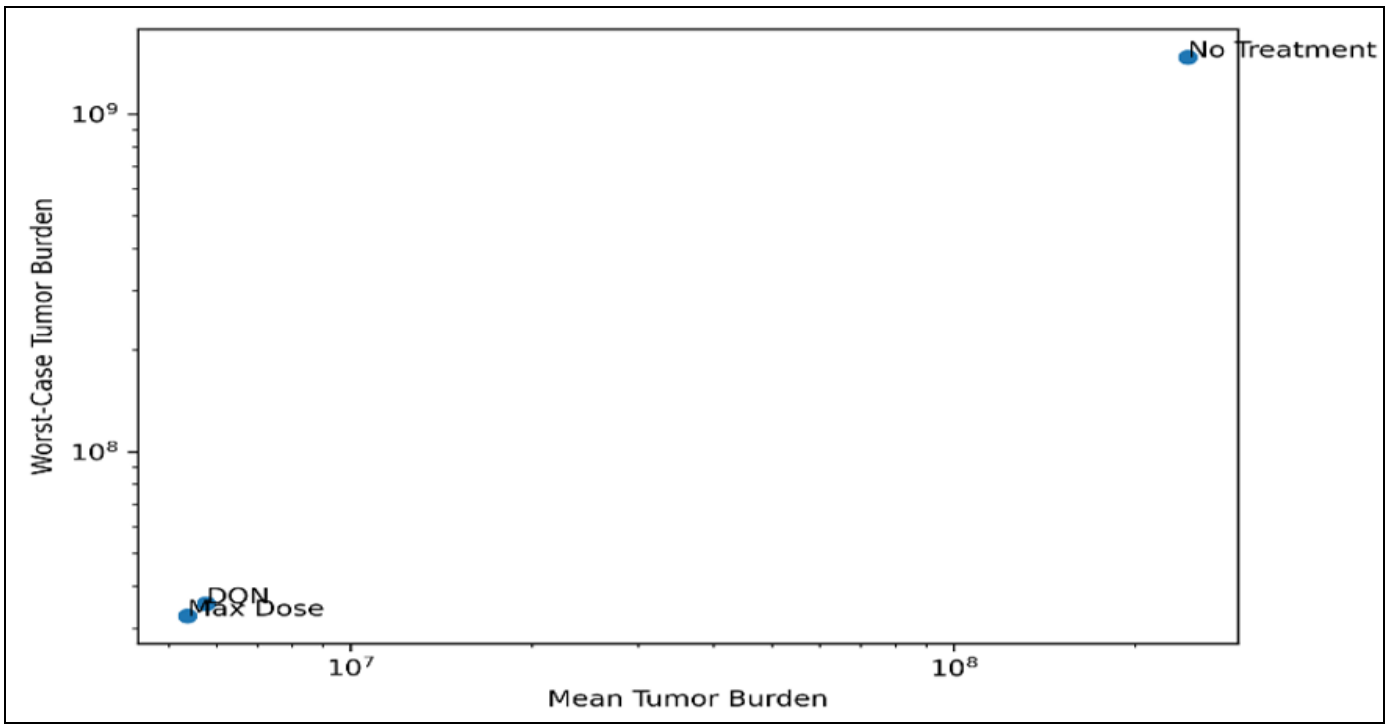

Fig 8 Mean Versus Worst-Case Tumor Burden Under Aggressive Conditions

Overall, the results highlight three key trade-offs. First, aggressiveness versus adaptivity: MTD achieves marginally superior tumor suppression at the cost of continuous high-intensity dosing, while DQN accepts a 5-30% increase in tumor burden to enable adaptive control. Second, expected performance versus tail risk: the DQN exhibits slightly higher CVaR but remains within 10-12% of MTD, indicating strong risk control. Third, optimality versus robustness: under distribution shift, the DQN shows less than 9% degradation relative to MTD, demonstrating resilience to unseen tumor dynamics. Together, these findings support the viability of reinforcement learning for adaptive cancer therapy when evaluated through a risk-aware lens.

From a clinical perspective, the results support the hypothesis that adaptive therapy can achieve competitive tumor control while avoiding continuous maximal dosing. The findings emphasize that reinforcement learning policies should be assessed as decision-support tools rather than black-box optimizers, and that risk-aware evaluation is essential for safety-critical applications. From a methodological standpoint, the study highlights the importance of reporting distributional metrics rather than single averages, stress-testing policies under adverse conditions, and designing evaluation frameworks that are reusable and extensible. These principles extend beyond oncology to other high-stakes healthcare domains.

Despite these contributions, several limitations remain. The tumor model is simplified and does not capture spatial heterogeneity, immune response, or multi-drug interactions. Chemotherapy dosing is discretized, whereas real-world dosing decisions are continuous and constrained by pharmacokinetics. All experiments are conducted in simulation, limiting immediate clinical applicability. Finally, the study focuses on a DQN policy, leaving other reinforcement learning paradigms unexplored. These limitations motivate future research.

V. CONCLUSION AND FUTURE WORK

This study addressed a critical gap in reinforcement learning research for adaptive cancer therapy: the absence of standardized, risk-aware, and uncertainty-sensitive evaluation frameworks. Rather than proposing a new control algorithm in isolation, the work focused on how reinforcement learning policies should be evaluated to support clinically meaningful and safety-aware conclusions.

A modular simulation-based evaluation framework was developed using a biologically grounded two-population tumor growth model. The framework supports heterogeneous patient sampling, controlled distribution shifts, and reproducible experimentation. Beyond expected performance metrics, the study incorporated worst-case outcomes and Conditional Value at Risk to characterize tail-risk behavior, an aspect largely absent from prior reinforcement learning studies in oncology. Policies were evaluated under both standard and aggressive tumor dynamics to assess robustness and generalization.

Empirical results demonstrate that a Deep Q Network policy achieves tumor suppression comparable to maximum tolerated dose therapy, reducing tumor burden by over 97%

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

relative to no treatment and matching the same order of magnitude as aggressive fixed-dose strategies. This performance is achieved without constant maximal dosing, highlighting the ability of adaptive policies to exploit tumor dynamics while potentially reducing cumulative treatment exposure. Although the learned policy exhibits slightly higher mean and worst-case outcomes than maximum-dose therapy, increases in Conditional Value at Risk remain consistently below 12% across all evaluation settings, indicating controlled risk exposure.

Under distribution shift, all policies experience performance degradation; however, the relative gap between the adaptive policy and maximum-dose therapy remains below 9%, demonstrating strong robustness to unseen tumor dynamics and parameter uncertainty. These findings reinforce the central insight of this work: expected performance alone is insufficient for assessing clinical viability. Risk-sensitive and distribution-aware evaluation is essential for responsible reinforcement learning in safety-critical healthcare domains.

Overall, this study demonstrates that reinforcement learning-based adaptive chemotherapy can achieve competitive tumor control while maintaining acceptable risk profiles and robustness to uncertainty. More importantly, it establishes that how reinforcement learning policies are evaluated is as critical as how they are trained. By introducing a clinically aligned, risk-aware evaluation framework, this work contributes a necessary methodological foundation for advancing reinforcement learning in oncology and beyond. The reported results are obtained under controlled simulation settings and do not imply direct clinical safety or efficacy in real-world practice.

Several directions for future research emerge from this work. Incorporating robust reinforcement learning or distributionally robust optimization techniques could explicitly optimize worst-case and CVaR objectives. Extending the framework to multi-objective reinforcement learning would enable joint optimization of tumor suppression, toxicity, quality of life, and treatment cost. Greater clinical realism could be achieved by adopting continuous dosing actions and integrating pharmacokinetic-pharmacodynamic models. Introducing partial observability and noisy tumor state measurements would allow investigation of reinforcement learning under realistic clinical uncertainty.

REFERENCES

[1]. A. Eastman, J. S. Brown, and J. J. Cunningham, “Reinforcement learning-derived chemotherapeutic schedules for robust patient-specific therapy,” Nature Machine Intelligence, vol. 3, no. 12, pp. 1091-1101, Dec. 2021. https://doi.org/10.1038/s42256-021-00424-3.

[2]. X. Fu, Y. Luo, and B. Schölkopf, “Off-policy evaluation in reinforcement learning: A survey,” ACM Computing Surveys, vol. 56, no. 1, Art. no. 9, pp. 1-39, 2024. https://doi.org/10.1145/3594560.

[3]. O. Gottesman et al., “Guidelines for reinforcement learning in healthcare,” Nature Medicine, vol. 25, no. 1, pp. 16-18, Jan. 2019. https://doi.org/10.1038/s41591-018-0310-5.

[4]. N. Jiang and L. Li, “Doubly robust off-policy value evaluation for reinforcement learning,” in *Proceedings of the 33rd International Conference on Machine Learning*, PMLR, 2016, pp. 652-661.

[5]. M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, “The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care,” Nature Medicine, vol. 24, no. 11, pp. 1716-1720, Nov. 2018. https://doi.org/10.1038/s41591-018-0213-5.

[6]. M. Labrie, M. Tannenbaum, and R. A. Gatenby, “Adaptive therapy in oncology: Principles and perspectives,” Cancer Research, vol. 82, no. 15, pp. 2761-2769, Aug. 2022. https://doi.org/10.1158/0008-5472.CAN-21-3849.

[7]. S. A. Murphy, “Optimal dynamic treatment regimes,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 65, no. 2, pp. 331-355, 2003. https://doi.org/10.1111/1467-9868.00389.

[8]. S. Nemati, M. M. Ghassemi, and G. D. Clifford, “Optimal medical therapy dosing from suboptimal clinical examples: A deep reinforcement learning approach,” in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2016, pp. 2978-2981. https://doi.org/10.1109/EMBC.2016.7591355.

[9]. S. Padmanabhan, A. Mesbah, and Y. Shen, “Reinforcement learning in healthcare: A systematic review,” Artificial Intelligence in Medicine, vol. 145, Art. no. 102657, 2024. https://doi.org/10.1016/j.artmed.2023.102657.

[10]. D. Precup, R. S. Sutton, and S. Singh, “Eligibility traces for off-policy policy evaluation,” in *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, 2000, pp. 759-766.

[11]. Y. Shen, Y. Wu, and Z. Wang, “Deep reinforcement learning for chemotherapy scheduling,” IEEE Transactions on Biomedical Engineering, vol. 64, no. 6, pp. 1387-1398, Jun. 2017. https://doi.org/10.1109/TBME.2016.2608790.

[12]. A. Singh, R. Kumar, and D. Gupta, “Sequential decision-making in oncology using reinforcement learning,” Frontiers in Oncology, vol. 14, Art. no. 1278456, 2024. https://doi.org/10.3389/fonc.2024.1278456.

[13]. X. Sun, Y. Zhang, and J. Li, “Adaptive cancer therapy via deep reinforcement learning,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 8, pp. 4021-4032, Aug. 2023. https://doi.org/10.1109/JBHI.2023.3279410.

[14]. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[15]. P. S. Thomas and E. Brunskill, “Data-efficient off-policy policy evaluation for reinforcement learning,”

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26jan1026

in *Proceedings of the 33rd International Conference on Machine Learning*, PMLR, 2016, pp. 2139-2148.

[16]. L. Xu, S. Zhu, and N. Wen, "Deep reinforcement learning and its applications in medical imaging and radiation therapy: A survey," Physics in Medicine and Biology, vol. 67, no. 22, Art. no. 22TR02, 2022. https://doi.org/10.1088/1361-6560/ac9cb3.

[17]. C. Yu, J. Liu, S. Nemati, and F. Doshi-Velez, "Reinforcement learning in healthcare: A survey," ACM Computing Surveys, vol. 55, no. 1, Art. no. 5, pp. 1-36, 2019. https://doi.org/10.1145/3312042.