ORIGINAL RESEARCH **Open access**

Deep Reinforcement Learning for Personalized Adaptive Radiation Therapy Planning in Head and Neck Cancer Using Daily Cone-Beam CT and Dosimetric Constraints

Andreas Müller[1*], Stefan Weber[1], Julia Hoffmann[2], Lukas Schneider[1], Tobias Klein[2]

Abstract

Head and neck cancer radiotherapy requires highly precise dose delivery to ensure tumor control while sparing nearby critical structures, but daily anatomical changes such as tumor shrinkage, weight loss, and setup variability often degrade treatment accuracy. Although cone-beam CT provides valuable daily imaging, current adaptive radiotherapy workflows remain largely manual, time-consuming, and infrequent, limiting their ability to respond to ongoing anatomical changes and often resulting in suboptimal target coverage or increased toxicity risk. To address these limitations, we propose a deep reinforcement learning framework for fully automated daily treatment adaptation using cone-beam CT and dosimetric constraints. The problem is formulated as a sequential decision-making task in which an agent adjusts beam parameters based on evolving patient anatomy, cumulative dose, and constraint satisfaction. The state includes daily imaging and dose history, the action space involves fluence or multileaf collimator adjustments, and the reward function balances target coverage, organ-at-risk sparing, and plan stability. A patient-specific simulator based on historical imaging enables training without real-time patient interaction. This framework enables continuous, personalized, and automated plan adaptation that directly responds to anatomical changes while maintaining clinical safety constraints. By leveraging long-horizon optimization, the system can outperform static planning strategies and better manage stochastic anatomical variations in head and neck cancer treatment. Overall, this approach provides a foundation for closed-loop adaptive radiotherapy that could improve treatment accuracy, reduce toxicity, and reduce reliance on manual planning.

Keywords Deep reinforcement learning, Adaptive radiation therapy, Head and neck cancer, Cone-beam computed tomography, Dosimetric constraints, Personalized treatment planning

*Correspondence:
Andreas Müller
andreas.mueller@gmail.com
[1] Department of Healthcare Data Analytics, Heidelberg University, Heidelberg, Germany
[2] Department of Intelligent Clinical Systems, Technical University of Munich, Munich, Germany

Introduction

Head and neck cancer radiotherapy serves as a primary curative or adjuvant modality for a wide spectrum of malignancies arising in the oral cavity, oropharynx, larynx, and hypopharynx. The complex anatomy of this region places multiple organs at risk in close proximity to target volumes, including the parotid glands, spinal cord, brainstem, mandible, and optic structures. Intensity-modulated techniques are routinely employed to sculpt high-dose regions around tumors while minimizing exposure to these critical structures, yet the inherent sensitivity of surrounding tissues demands meticulous planning to balance efficacy and safety [1, 2].

Anatomical changes during the typical six- to seven-week treatment course cause the delivered dose to deviate substantially from the original plan. Tumor shrinkage on the order of one to two millimeters per week, combined with patient weight loss and daily setup variations, displaces both targets and organs at risk in unpredictable ways. These deformations accumulate over fractions, leading to underdosing of planning target volumes or unintended overdosing of organs at risk that were contoured on the initial simulation scan [3, 4].

Current adaptive radiotherapy strategies remain reactive, labor-intensive, and limited in frequency, typically performed only once or twice during the entire course. Re-planning requires acquisition of a new computed tomography scan, manual re-contouring of all structures, and full re-optimization of beam parameters, processes that consume hours of clinician effort. As a result, most treatment fractions proceed without adaptation, accepting dosimetric compromises that could otherwise be mitigated through more frequent intervention [5, 6].

This work proposes deep reinforcement learning as the core engine for daily, automated, and personalized plan adaptation that directly incorporates daily cone-beam computed tomography and explicit dosimetric constraints. The framework treats each treatment fraction as a sequential decision step within a Markov decision process, enabling the agent to learn policies that optimize fluence adjustments in response to observed anatomical changes. This roadmap outlines the complete conceptual architecture, from state-action formulation to safety mechanisms, establishing a foundation for fully autonomous adaptive radiotherapy in head and neck cancer [7, 8].

Figure 1 illustrates the full conceptual architecture through which daily cone-beam CT, cumulative dosimetry, constrained Markov decision-making, deep reinforcement learning policy optimization, safety enforcement, and clinician-supervised delivery are integrated into a personalized adaptive radiotherapy workflow for head and neck cancer.

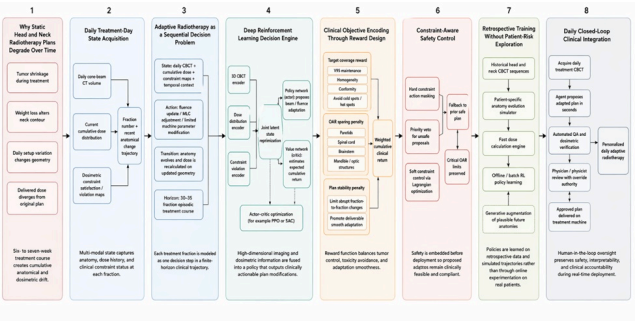

Figure 1. Conceptual architecture of deep reinforcement learning for daily personalized adaptive radiation therapy in head and neck cancer using cone-beam CT and dosimetric constraints.

Background

Head and neck radiotherapy planning

Intensity-modulated radiation therapy and volumetric modulated arc therapy constitute the standard of care for head and neck cancer, allowing highly conformal dose delivery to gross tumor volume, clinical target volume, and planning target volume expansions. Target volumes are defined according to established guidelines that account for microscopic disease spread, while organs at risk receive strict dose-volume constraints to prevent severe toxicity. Typical constraints include a mean dose below 26 Gy to the parotid glands to preserve salivary function, a maximum dose below 45 Gy to the spinal cord, and below 54 Gy to the brainstem, with additional limits applied to the mandible and optic structures to avoid osteoradionecrosis or vision loss [1, 9].

Treatment planning systems solve inverse optimization problems to satisfy these competing objectives, yet the resulting plans remain static and unable to accommodate daily anatomical variations. Planners must manually balance trade-offs between target coverage and organ-at-risk sparing through iterative adjustments of beam weights and multileaf collimator sequences. The complexity of head and neck anatomy, combined with the need for simultaneous integrated boost regimens, further increases the computational and clinical burden of generating high-quality plans that remain robust over the full treatment course [2, 10].

Anatomical changes during treatment

Tumor shrinkage occurs at a rate of approximately one to two millimeters per week in responding head and neck lesions, while concurrent weight loss induces medial migration of the parotid glands and alterations in neck contour. These changes shift the relative geometry between targets and organs at risk, causing the high-dose gradient regions of intensity-modulated plans to migrate into previously spared structures. Setup errors on the order of several millimeters compound the problem, as daily positioning inaccuracies further distort the delivered dose distribution relative to the simulation geometry [4, 11].

The dosimetric impact of these anatomical deformations can be profound, with studies demonstrating systematic increases in parotid mean dose and spinal cord maximum dose when adaptation is not performed. Weight loss alone can increase mean parotid dose by several gray over the treatment course, elevating the risk of xerostomia. Without daily monitoring and correction, the cumulative effect across thirty to thirty-five fractions leads to clinically relevant deviations that undermine both tumor control probability and normal tissue complication probability [5, 12].

Current adaptive radiotherapy

Adaptive radiotherapy for head and neck cancer typically relies on threshold-based triggers, such as observed anatomical changes exceeding five to ten percent on weekly imaging. Once triggered, the workflow involves acquisition of a new computed tomography scan, re-contouring of all target and organ-at-risk volumes, and complete re-optimization of the treatment plan using commercial treatment planning systems. This offline process is performed at most once or twice per course, leaving the majority of fractions delivered under the original plan [3, 6].

Limitations of current adaptive workflows stem from their labor-intensive nature and inability to scale to daily use. Each adaptation cycle requires hours of physician and physicist time, creating bottlenecks in busy clinics and restricting application to only the most pronounced anatomical shifts. Moreover, the discrete, infrequent nature of re-planning fails to capture the continuous trajectory of patient-specific changes, resulting in suboptimal cumulative dose distributions that do not fully exploit the potential of image-guided delivery [2, 5].

Reinforcement learning in medical physics

Reinforcement learning has demonstrated utility in medical physics for sequential decision tasks such as beam angle selection, fluence map optimization, and motion management across various disease sites. Early applications formulated treatment planning as a Markov decision process in which an agent learns to adjust beam parameters to maximize a dosimetric reward while satisfying clinical constraints. These approaches have shown promise in automating inverse planning and adapting to dynamic conditions without requiring exhaustive search of the solution space [7, 8].

Deep reinforcement learning extends classical methods by employing neural network function approximators to handle high-dimensional state spaces, such as volumetric imaging data. In radiation therapy contexts, actor-critic architectures have been explored for automated plan adaptation in lung and prostate cancer, providing a foundation for extension to head and neck sites. Prior work highlights the capacity of deep reinforcement learning to discover policies that balance competing objectives and generalize across patient anatomies when trained on appropriate simulators [13-15].

Markov Decision Process (mdp) Formulation

State space

The state space integrates daily cone-beam computed tomography volumes as the primary imaging input, providing three-dimensional anatomical information at the time of each fraction. Additional state components include the current cumulative dose distribution warped onto the daily anatomy and a vector encoding the fraction number along with recent anatomical change metrics derived from deformable registration. Explicit dosimetric constraint satisfaction maps, represented as three-dimensional tensors highlighting voxels exceeding organ-at-risk limits, complete the state representation to ensure the agent has full awareness of clinical objectives [4, 9].

This multi-modal state design enables the agent to perceive both geometric deformations and their dosimetric consequences in a single observation. By incorporating temporal elements such as change trajectory, the state

captures the patient-specific evolution pattern rather than treating each fraction in isolation. The resulting high-dimensional but information-rich state supports deep neural network processing while remaining computationally tractable through appropriate feature extraction [11, 16].

Action space

The action space comprises continuous or discretized updates to beam intensity weights and fluence maps for each treatment field, allowing fine-grained adjustments to the dose distribution. Multileaf collimator leaf positions may also be included as actions to enable direct modification of aperture shapes when full re-optimization is not required. Limited gantry angle adjustments can be incorporated for arc therapies, providing additional degrees of freedom while respecting mechanical constraints of the linear accelerator [17, 18].

Continuous action representations are preferred to capture the nuanced trade-offs inherent in head and neck planning, where small fluence perturbations can substantially improve organ-at-risk sparing without compromising target coverage. The action space is bounded to ensure feasible modifications that maintain deliverability within clinical time slots. Parameterization through policy networks allows the agent to output adjustments that are directly translatable to machine parameters via the treatment planning system interface [19, 20].

Transition dynamics

Transition dynamics model the stochastic evolution of patient anatomy between fractions, driven by tumor response, weight loss, and random setup variations that cannot be fully predicted. The effect of each selected action is simulated by recalculating the dose distribution on the updated anatomy using a fast dose engine, yielding the next state and associated reward. Patient-specific simulators are constructed from historical cone-beam computed tomography sequences to approximate real-world dynamics during offline training [21, 22].

Because exact future anatomy remains unknown, the transition function incorporates probabilistic elements derived from population statistics or generative models trained on prior head and neck cohorts. The simulator must faithfully reproduce both geometric deformations and their impact on dose deposition to enable the agent to learn robust policies. This offline simulation paradigm avoids any requirement for online exploration on actual patients, preserving safety throughout the learning process [23, 24].

Horizon and discounting

The decision process is formulated as a finite-horizon episodic task spanning the thirty to thirty-five fractions of a standard head and neck regimen, with each fraction constituting a single time step. A discount factor in the range of 0.95 to 0.99 is applied to prioritize near-term dosimetric improvements while still valuing long-term cumulative outcomes at treatment completion. The terminal state is reached upon delivery of the final fraction, at which point the episode reward reflects the overall plan quality across the entire course [14, 25].

Discounting encourages the agent to balance immediate constraint satisfaction with sustained performance over the full horizon, preventing myopic policies that sacrifice later fractions for short-term gains. Episodic termination aligns naturally with clinical treatment completion, allowing the value function to estimate expected future returns from any intermediate state. This formulation ensures that learned policies remain clinically relevant by optimizing the complete patient journey rather than isolated daily decisions [10, 26].

Table 1 formalizes the proposed framework by mapping the clinical adaptive radiotherapy problem onto explicit Markov decision process components and their head and neck cancer-specific interpretations.

Table 1. Structural mapping of the adaptive radiotherapy problem to a constrained Markov decision process for head and neck cancer.

MDP Component	Framework Definition in This Manuscript	Head and Neck Cancer-Specific Instantiation	Clinical Significan
Agent	Deep reinforcement learning policy that selects plan adaptations at each fraction	Actor-critic policy operating on treatment-day imaging and dosimetry	Replaces infrequen manual re planning wi consisten daily decisi support

Environment	Dynamic radiotherapy treatment course with evolving anatomy and cumulative dose effects	Daily anatomical change driven by tumor shrinkage, weight loss, and setup variation	Determine whether a static plan becomes progressive suboptima
State	Multi-modal observation available before each fraction	Daily CBCT, cumulative dose, constraint violation maps, fraction index, anatomical change trajectory	Captures bo geometry a delivered treatment history
Action	Deliverable adaptation selected by the policy	Fluence updates, beam weight perturbations, MLC adjustments, limited machine-feasible parameter changes	Converts anatomica awareness i actionable planning decisions
Transition Function	Evolution from current fraction to next fraction after action application	Dose recalculation on updated anatomy plus stochastic patient evolution between fractions	Links prese adaptation choices to la cumulative dosimetric outcomes
Reward	Scalar objective guiding policy optimization	Weighted combination of target coverage gains, OAR penalties, and stability constraints	Encodes th true clinica trade-off structure o radiotherap adaptation
Policy	Decision rule mapping state to action	Personalized daily adaptive strategy conditioned on anatomy and dose history	Enables patient-spec rather tha protocol-fix adaptation
Value Function	Expected future cumulative return from a given state	Forecast of downstream dosimetric consequences across remaining fractions	Prevents myopic fraction-by fraction optimizatio
Horizon	Finite episodic treatment sequence	Approximately 30–35 fractions in a standard head and neck regimen	Makes adaptation inherently longitudina rather tha isolated
Discounting	Controlled weighting of immediate versus future outcomes	Near-term dosimetric corrections balanced against end-of-course cumulative quality	Reflects th clinical need preserve bo immediate safety and long-term tumor contr
Constraint Structure	Explicit safety boundaries superimposed on the MDP	Hard serial-organ limits and soft multi-objective trade-offs for parallel organs	Prevents clinically unacceptab explorator behavior
Terminal Outcome	Final treatment-course result after last fraction	End-of-course target coverage, OAR exposure, and cumulative plan quality	Determine whether personalizat improves th full regime

Deep RL Architecture

CBCT encoder

A three-dimensional convolutional neural network serves as the CBCT encoder, extracting hierarchical spatial features from the daily volumetric images while preserving critical anatomical context. Pre-training on large radiotherapy computed tomography and cone-beam computed tomography datasets enables the encoder to capture clinically relevant patterns such as tumor boundaries and organ-at-risk contours without requiring manual annotation at inference time. Dimensionality reduction through pooling and bottleneck layers produces a compact latent representation suitable for fusion with other state components [7, 23].

Residual connections and attention mechanisms within the encoder further enhance the model's ability to focus on regions of high dosimetric importance, such as areas near the spinal cord or parotid glands. The architecture is designed to handle the lower soft-tissue contrast typical of cone-beam computed tomography by incorporating domain-adaptation layers that align features with simulation computed tomography distributions. This robust encoding step ensures that anatomical changes are faithfully represented in the policy input regardless of daily image quality variations [15, 16].

Dose and constraint encoder

The current dose distribution is processed as a three-dimensional tensor in parallel with the CBCT encoder, allowing the network to correlate spatial anatomy with accumulated dose deposition. Constraint violation maps are generated by thresholding the dose volume against clinical organ-at-risk limits and concatenated as additional input channels, providing an explicit signal of safety status. Shared convolutional layers fuse dose and constraint features into a unified embedding that highlights regions requiring immediate attention [8, 18].

This encoder design enables the agent to reason jointly about geometry and dosimetry, a capability essential for head and neck adaptation where small anatomical shifts can produce large dosimetric consequences. Batch normalization and skip connections stabilize training across heterogeneous patient anatomies and varying dose accumulation stages. The resulting joint representation feeds directly into the policy and value heads, ensuring that all actions are informed by both imaging and dosimetric context [20, 22].

Policy and value networks

An actor-critic architecture is employed to handle the continuous action space inherent in fluence map optimization, with proximal policy optimization or soft actor-critic serving as the base algorithm. The policy network outputs a Gaussian distribution over action perturbations, while the value network estimates expected future returns to guide advantage computation and reduce variance during updates. Both networks share the encoder backbones to promote efficient feature reuse across actor and critic pathways [13, 14].

Separate heads for mean and standard deviation parameters in the policy allow controlled exploration during training while maintaining deterministic behavior at deployment. Entropy regularization terms encourage sufficient exploration of the action space without destabilizing convergence on safety-critical tasks. The overall architecture is optimized end-to-end using clipped surrogate objectives, yielding stable policies suitable for clinical translation [19, 26].

Reward Design (CRITICAL)

Target coverage rewards

Positive rewards are assigned when planning target volume coverage metrics satisfy clinical thresholds, such as V95 percent exceeding 99 percent and V107 percent remaining below 2 percent across all target structures. The reward component scales continuously with improvements in homogeneity and conformity indices, providing dense feedback that guides the agent toward clinically acceptable target dosing. Cold spots or hot spots within the planning target volume incur graduated negative rewards proportional to their dosimetric severity, ensuring the agent prioritizes uniform coverage [10, 25].

This target-centric reward structure aligns directly with the primary goal of radiotherapy, which is to deliver adequate dose to malignant tissue while avoiding underdosing that could compromise local control. By incorporating both binary threshold satisfaction and continuous metric gradients, the reward function supports fine-grained learning of nuanced trade-offs. Such design prevents the

agent from exploiting loopholes that achieve marginal coverage at the expense of overall plan quality [7, 16].

OAR sparing penalties

Negative rewards are levied for any violation of organ-at-risk dose constraints, with penalties weighted according to clinical severity such that spinal cord and brainstem exceedances incur substantially larger costs than parotid mean dose violations. The penalty magnitude scales with both the volume of violation and the degree of overdose, creating a strong gradient that discourages actions leading to critical structure toxicity. Mean and maximum dose metrics for each organ at risk are evaluated independently to provide comprehensive safety feedback [1, 2].

This hierarchical penalty scheme reflects established radiation oncology priorities, where serial organs like the spinal cord demand absolute protection while parallel organs like the parotids tolerate moderate dose increases. The reward design thus embeds domain knowledge directly into the learning objective, enabling the agent to internalize clinical trade-off preferences without explicit programming. Weighted summation across all organs at risk ensures balanced sparing that mirrors multi-objective clinical planning [14, 20].

Stability penalty

A dedicated stability penalty discourages large fraction-to-fraction changes in beam parameters or fluence maps, promoting smooth adaptation trajectories that maintain plan deliverability and reduce mechanical stress on the linear accelerator. The penalty term is proportional to the L2 norm of action differences relative to the previous fraction, with a tunable coefficient that balances adaptation aggressiveness against plan consistency. This component prevents oscillatory behavior that could arise from over-reactive responses to minor anatomical fluctuations [15, 22].

Incorporating stability into the reward encourages policies that evolve gradually across the treatment course, aligning with the gradual nature of anatomical changes in head and neck cancer. The penalty also facilitates clinical acceptance by producing adaptation sequences that physicians can review and approve without abrupt shifts from one day to the next. Overall, the stability term contributes to safer, more predictable treatment delivery while preserving the benefits of daily personalization [19, 24].

Safety Constraints

Hard constraints

Hard constraints are enforced through action masking that immediately invalidates any proposed fluence adjustment or multileaf collimator position capable of exceeding critical organ-at-risk limits such as the spinal cord maximum dose of 45 Gy or the brainstem maximum of 54 Gy. This deterministic filtering ensures that the policy network never outputs unsafe actions even during exploratory phases of training, preserving patient safety as a non-negotiable boundary condition. The masking mechanism operates on the constraint satisfaction maps embedded in the state, allowing the agent to focus exploration exclusively within the feasible region defined by clinical guidelines [1, 2].

Critical structure protection is further strengthened by priority-based veto layers that override the policy output whenever a hard limit violation is detected in the forward dose calculation. Such vetoes trigger an immediate fallback to the previous fraction’s plan, maintaining treatment continuity without interruption. By embedding these safeguards directly into the architecture, the framework guarantees that daily adaptations remain strictly compliant with established safety thresholds regardless of anatomical complexity or stochastic variations [9, 20].

Soft constraints with lagrangian

Soft constraints are incorporated via Lagrangian relaxation within a constrained reinforcement learning formulation, where a separate multiplier dynamically balances the primary reward objective against cumulative organ-at-risk violation penalties. Primal-dual optimization updates the multiplier online during training to enforce long-term satisfaction of mean and maximum dose limits without requiring perfect hard enforcement on every step. This approach permits controlled, temporary relaxations when clinically justified by target coverage needs while still converging toward feasible policies over the full treatment horizon [10, 14].

The Lagrangian formulation enables explicit trade-off tuning between competing objectives, allowing the agent to learn nuanced behaviors that mirror radiation oncologist decision-making under uncertainty. Dual variables are adjusted based on observed violation frequency across simulated trajectories, ensuring the policy internalizes clinical priorities without manual weighting. Consequently, the resulting adaptive strategies achieve superior

dosimetric balance compared to purely unconstrained optimization while retaining flexibility for patient-specific anatomical evolution [26, 27].

Training and Simulation

Training environment

The training environment is constructed as a high-fidelity patient-specific simulator that replays sequences of historical cone-beam computed tomography images acquired at multiple time points during prior head and neck treatments. A generative model augments these sequences to produce plausible future anatomies that capture tumor shrinkage, weight loss, and setup variations, thereby exposing the agent to a diverse range of realistic transition dynamics. Fast dose engines embedded within the simulator compute the immediate dosimetric consequences of each action on the warped anatomy, closing the observation-reward loop required for reinforcement learning [21, 22].

Historical datasets provide the foundation for simulator fidelity, ensuring that learned policies generalize across the anatomical variability observed in real clinical cohorts. The environment supports parallel rollouts across thousands of virtual patients, accelerating convergence while maintaining computational efficiency suitable for large-scale deep network training. By grounding the simulation exclusively in retrospective cone-beam computed tomography data, the framework avoids any reliance on idealized assumptions and instead mirrors the stochastic, patient-driven changes encountered in daily practice [23, 24].

Offline RL (Batch RL)

Offline reinforcement learning is adopted to eliminate the need for online exploration on actual patients, relying instead on large batches of pre-collected state-action-reward trajectories generated from the simulator. Conservative Q-learning variants are employed to prevent overestimation of out-of-distribution actions, ensuring that the policy remains safe and aligned with demonstrated clinical behaviors. This batch-oriented approach further mitigates distributional shift by constraining updates to actions already present in the historical dataset, thereby supporting reliable deployment without real-time risk [13, 15].

Quantum-inspired extensions of deep reinforcement learning have been explored in related clinical decision support contexts to accelerate convergence on high-dimensional radiotherapy problems, offering potential efficiency gains for future large-scale training. The offline paradigm aligns naturally with regulatory requirements for medical software, as all policy learning occurs on retrospective data before any prospective evaluation. Consequently, the framework maintains strict separation between training and clinical use, preserving the integrity of patient safety protocols throughout development [27, 28].

Clinical Integration

Daily workflow

The daily workflow begins with acquisition of the treatment-day cone-beam computed tomography, which is automatically fed into the trained deep reinforcement learning agent for rapid inference of fluence map adjustments. The agent outputs an updated plan proposal within seconds, complete with predicted dose distribution and constraint satisfaction metrics, enabling seamless integration into existing linear accelerator consoles. A physicist performs automated quality assurance checks before the proposal advances to physician review, ensuring that only clinically viable adaptations reach the approval stage [3, 6].

Upon physician approval, the updated parameters are transferred directly to the treatment machine for immediate delivery, completing the closed-loop adaptation cycle within the standard time slot allocated for image-guided radiotherapy. The entire process is orchestrated through a dedicated middleware layer that interfaces with commercial treatment planning and record-and-verify systems. This streamlined sequence transforms daily cone-beam computed tomography from a verification tool into an active driver of personalized treatment, minimizing workflow disruption while maximizing dosimetric fidelity [28, 29].

Human-in-the-loop

Human-in-the-loop oversight is preserved through an interactive interface that presents the agent’s proposed adaptation alongside side-by-side visualizations of the original plan, current cumulative dose, and projected organ-at-risk metrics. The attending radiation oncologist retains full override authority, allowing manual adjustment or rejection of the suggestion based on additional clinical

context not captured in the state representation. Safety monitoring dashboards continuously track key performance indicators across fractions, triggering alerts if any deviation from expected trajectories is detected [5, 12].

This collaborative design fosters gradual clinical acceptance by positioning the reinforcement learning agent as a decision-support tool rather than an autonomous replacement for human judgment. Threshold-based approval rules can be configured to require mandatory review for high-risk adaptations, such as those near serial organs, while permitting fully automated execution for routine cases. Over time, accumulated override data can be incorporated into retraining cycles to further refine the policy toward physician-preferred behaviors [2, 15].

Evaluation Strategy

Dosimetric metrics

Evaluation relies on standard dosimetric metrics that quantify planning target volume coverage through V95 percent and V107 percent, homogeneity index, and conformity index computed on the daily anatomy. Organ-at-risk endpoints include mean dose to bilateral parotids, maximum point dose to the spinal cord and brainstem, and volume-based constraints for the mandible and optic structures, all accumulated across the full simulated course. These metrics are reported both per fraction and as cumulative totals to capture the longitudinal benefit of daily adaptation over static planning [1, 10].

Additional composite indices such as the generalized equivalent uniform dose and normal tissue complication probability models provide clinically interpretable summaries of plan quality. Comparisons are performed against both the original non-adaptive plan and conventional offline adaptive schedules to quantify incremental improvements attributable to the reinforcement learning policy. The metric suite is deliberately aligned with international reporting guidelines to facilitate direct translation into multi-center validation studies [9, 20].

Validation protocols

Validation protocols center on retrospective evaluation using held-out historical patient datasets that include complete cone-beam computed tomography sequences and delivered dose reconstructions. The deep reinforcement learning agent is applied in simulation mode to generate daily adapted plans, which are then benchmarked against the actual clinically delivered plans and against standard non-daily adaptive strategies. Statistical analysis employs paired Wilcoxon tests and dose-volume histogram comparisons to establish superiority in target coverage and organ-at-risk sparing under identical anatomical conditions [21, 22].

Cross-validation across multiple institutions ensures generalizability, with separate cohorts reserved for hyperparameter tuning and final testing. Simulator-based stress testing introduces controlled perturbations in anatomy evolution rates to assess policy robustness under varying clinical scenarios. These protocols collectively demonstrate the framework's readiness for prospective trials while satisfying the evidentiary requirements for regulatory clearance of adaptive radiotherapy software [28, 29].

Table 2 clarifies the analytical advantage of the proposed approach by contrasting conventional adaptive radiotherapy workflows with a deep reinforcement learning-driven daily personalization paradigm.

Table 2. Comparative analytical framework distinguishing conventional adaptive radiotherapy from deep reinforcement learning-driven daily personalization.

Analytical Dimension	Conventional Static / Trigger-Based Adaptive Radiotherapy	Proposed Deep RL Daily Adaptive Framework	Why Differ Matt
Temporal logic	Reactive and intermittent	Proactive and fraction-by-fraction	The pro frame trea adaptati contin clinical p rather t occas corre
Decision structure	Human-initiated threshold response	Sequential policy-driven optimization	This s adaptati episo interver longitu

			decis mak
Use of daily CBCT	Primarily verification or trigger detection	Core state input for immediate action selection	Daily in beco operati decisive than m observa
Personalization level	Limited to major observed changes	Continuous personalization to each day's anatomy and dose history	Support grai correct patient-s anato evolu
Representation of prior treatment history	Often weakly incorporated	Explicitly encoded through cumulative dose and temporal context	Mak adapt aware c has al been de not just seen t
Optimization target	Single re-plan quality at selected time points	Total cumulative course quality across all fractions	Align optimiz objectiv the r longitu structu radioth
OAR protection mechanism	Planner-mediated trade-off during manual re-optimization	Reward penalties plus hard and soft constraint enforcement	Embeds safety directly comput deci proc
Handling of uncertainty	Mostly clinician judgment and offline reassessment	Simulator-trained policy acting under stochastic transition dynamics	Better m unpred anato change and r treatr
Labor burden	High, with repeated contouring and re-optimization	Reduced through automated plan proposal generation	Address of the barr preventi adapta prac
Speed of adaptation	Hours to days	Seconds to minutes for inference plus verification	Makes day or slot ada operati plaus
Plan consistency	Variable across manual sessions	Stability-aware through explicit penalty terms	Helps erra adaptati impro deliver
Safety governance	Human review after plan generation	Safety embedded before and after plan generation	Produ stron assur architec clini deploy
Role of clinician	Primary optimizer and decision-maker	Supervisory expert with override authority	Prese physi control redu repet techn burd
Training paradigm	No learning across patient trajectories in workflow itself	Offline reinforcement learning on retrospective simulated trajectories	Allows improv bef deploy without pati experim
Scalability	Limited by staffing and workflow bottlenecks	Potentially scalable across daily fractions and larger patient volumes	Daily ad radioth become realis routine p
Conceptual contribution	Improved re-planning procedure	Closed-loop adaptive treatment intelligence	Th manus novelty refra planni safe sec clinical

Conclusion

The proposed deep reinforcement learning framework establishes a comprehensive conceptual architecture for daily, personalized adaptive radiation therapy planning in head and neck cancer that directly leverages cone-beam computed tomography and explicit dosimetric constraints. By formulating the problem as a Markov decision process with carefully designed state, action, and reward components, the system enables sequential decision-making that continuously optimizes treatment delivery in response to patient-specific anatomical changes. The integration of actor-critic networks, safety mechanisms, and offline training paradigms provides a complete blueprint for automated adaptation that maintains clinical objectives across the entire treatment course.

Key advantages include fully personalized plan updates that respond to daily imaging, automated operation that reduces labor burden, and explicit awareness of organ-at-risk constraints that safeguards critical structures such as the spinal cord and parotids. The framework further promotes plan stability and human oversight, facilitating smooth clinical translation while preserving physician authority. These elements collectively position deep reinforcement learning as a transformative tool for achieving closed-loop adaptive radiotherapy that was previously unattainable with conventional methods.

Limitations of the current conceptual design include dependence on high-fidelity patient-specific simulators for training, the need for rigorous safety certification before prospective deployment, and the requirement for sustained physician acceptance through transparent human-in-the-loop interfaces. Additional challenges arise from the computational demands of three-dimensional convolutional processing and the necessity to validate generalization across diverse head and neck subsites and treatment regimens. Addressing these limitations through continued simulator refinement and multi-institutional data sharing will be essential for successful clinical adoption.

Future work should prioritize implementation of the framework on large historical head and neck datasets and seamless integration into commercial treatment planning systems to enable prospective evaluation. Collaborative efforts between artificial intelligence researchers, medical physicists, and radiation oncologists will accelerate the transition from conceptual design to routine clinical use. Ultimately, this reinforcement learning approach offers a pathway to safer, more effective, and truly personalized radiotherapy for head and neck cancer patients worldwide.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 Mar 2023 Revised: 26 May 2023 Accepted: 26 Jun 2023
Published online: 20 January 2024

References

Morgan HE, Sher DJ. Adaptive radiotherapy for head and neck cancer. Cancers Head Neck. 2020;5(1):1. https://doi.org/10.1186/s41199-019-0051-6.

Castelli J, Simon A, Lafond C, Perichon N, Rigaud B, Chajon E, et al. Adaptive radiotherapy for head and neck cancer. Acta Oncol. 2018;57(10):1284-92. https://doi.org/10.1080/0284186X.2018.1498154.

Yoon SW, Lin H, Alonso-Basanta M, Anderson N, Apinorasethkul O, Cooper K, et al. Initial evaluation of a novel cone-beam CT-based semi-automated online adaptive radiotherapy system for head and neck cancer treatment: a timing and automation quality study. Cureus. 2020;12(8):e9680. https://doi.org/10.7759/cureus.9680.

Hvid CA, Elstrøm UV, Jensen K, Grau C. Cone-beam computed tomography (CBCT) for adaptive image guided head and neck radiation therapy. Acta Oncol. 2018;57(4):552-6. https://doi.org/10.1080/0284186X.2017.1400682.

Håkansson K, Giannoulis E, Lindegaard A, Friborg J, Vogelius I. CBCT-based online adaptive radiotherapy for head and neck cancer: dosimetric evaluation of first clinical experience. Acta Oncol. 2023;62(11):1369-74. https://doi.org/10.1080/0284186X.2023.2263627.

Nasser N, Yang GQ, Koo J, Bowers M, Greco K, Feygelman V, et al. A head and neck treatment planning strategy for a CBCT-guided ring-gantry online adaptive radiotherapy system. J Appl Clin Med Phys. 2023;24(12):e14134. https://doi.org/10.1002/acm2.14134.

Tseng HH, Luo Y, Cui S, Chien JT, Ten Haken RK, Naqa IE. Deep reinforcement learning for automated radiation adaptation in lung cancer. Med Phys. 2017;44(12):6690-705. https://doi.org/10.1002/mp.12625.

Shen C, Gonzalez Y, Klages P, Qin N, Jung H, Chen L, et al. Intelligent inverse treatment planning via deep reinforcement learning: a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. Phys Med Biol. 2019;64(11):115013.

Vickress JR, Battista J, Barnett R, Yartsev S. Online daily assessment of dose change in head and neck radiotherapy without dose-recalculation. J Appl Clin Med Phys. 2018;19(5):659-65. https://doi.org/10.1002/acm2.12418.

Alfouzan AF. Radiation therapy in head and neck cancer. Saudi Med J. 2021;42(3):247-54. https://doi.org/10.15537/smj.2021.42.3.20200834.

Blumenfeld PA, Feldman J, Arbit E, Weizman N, Berger A, Hillman Y, et al. Daily artificial intelligence-assisted adaptive radiotherapy on cone-beam CT for cancer of the head and neck. Int J Radiat Oncol Biol Phys. 2022;114(3 Suppl):e590-1.

Blumenfeld PA, Arbit E, Wygoda MR, Berger A, Hillman Y, Weizman N, et al. Initial experience using daily artificial intelligence-assisted adaptive radiotherapy on cone-beam CT for bladder cancer. Int J Radiat Oncol Biol Phys. 2022;114(3 Suppl):e203-4.

Shen C, Nguyen D, Chen L, Gonzalez Y, McBeth R, Qin N, et al. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. Med Phys. 2020;47(6):2329-36. https://doi.org/10.1002/mp.14114.

Ebrahimi S, Lim GJ. A reinforcement learning approach for finding optimal policy of adaptive radiation therapy considering uncertain tumor biological response. Artif Intell Med. 2021;121:102193. https://doi.org/10.1016/j.artmed.2021.102193.

Sprouts D, Gao Y, Wang C, Jia X, Shen C, Chi Y. The development of a deep reinforcement learning network for dose-volume-constrained treatment planning in prostate cancer intensity modulated radiotherapy. Biomed Phys Eng Express. 2022;8(4):045008.

Leemans CR, Snijders PJF, Brakenhoff RH. The molecular landscape of head and neck cancer. Nat Rev Cancer. 2018;18(5):269-82. https://doi.org/10.1038/nrc.2018.11.

Hrinivich WT, Lee J. Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. Med Phys. 2020;47(12):6140-50. https://doi.org/10.1002/mp.14539.

Wang H, Bai X, Wang Y, Lu Y, Wang B. An integrated solution of deep reinforcement learning for automatic IMRT treatment planning in non-small-cell lung cancer. Front Oncol. 2023;13:1124458. https://doi.org/10.3389/fonc.2023.1124458.

Jones S, Thompson K, Porter B, Shepherd M, Sapkaroski D, Grimshaw A, et al. Automation and artificial intelligence in radiation therapy treatment planning. J Med Radiat Sci. 2024;71(2):290-8. https://doi.org/10.1002/jmrs.729.

Vlacich G, Stavas MJ, Pendyala P, Chen SC, Shyr Y, Cmelak AJ. A comparative analysis between sequential boost and integrated boost intensity-modulated radiation therapy with concurrent chemotherapy for locally advanced head and neck

cancer. Radiat Oncol. 2017;12(1):13. https://doi.org/10.1186/s13014-016-0759-1.

Avkshtol V, Meng B, Shen C, Choi BS, Okoroafor C, Moon D, et al. Early experience of online adaptive radiation therapy for definitive radiation of patients with head and neck cancer. Adv Radiat Oncol. 2023;8(5):101256. https://doi.org/10.1016/j.adro.2023.101256.

Mody MD, Rocco JW, Yom SS, Haddad RI, Saba NF. Head and neck cancer. Lancet. 2021;398(10318):2289-99. https://doi.org/10.1016/S0140-6736(21)01550-6.

Xu L, Zhu S, Wen N. Deep reinforcement learning and its applications in medical imaging and radiation therapy: a survey. Phys Med Biol. 2022;67(22):22TR02.

Taasti VT, Klages P, Parodi K, Muren LP. Developments in deep learning based corrections of cone beam computed tomography to enable dose calculations for adaptive radiotherapy. Phys Imaging Radiat Oncol. 2020;15:77-9. https://doi.org/10.1016/j.phro.2020.08.003.

Harjai N. Adaptive Radiation Therapy in Head and Neck Cancer.

Care P. Radiation therapy. Qual Assur. 2019;10(4).

Niraula D, Jamaluddin J, Matuszak MM, Ten Haken RK, Naqa IE. Quantum deep reinforcement learning for clinical decision support in oncology: application to adaptive radiotherapy. Sci Rep. 2021;11(1):23545. https://doi.org/10.1038/s41598-021-02851-7.

Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online adaptive radiation therapy. Int J Radiat Oncol Biol Phys. 2017;99(4):994-1003. https://doi.org/10.1016/j.ijrobp.2017.04.023.

van de Schoot AJ, de Boer P, Visser J, Stalpers LJA, Rasch CRN, Bel A. Dosimetric advantages of a clinical daily adaptive plan selection strategy compared with a non-adaptive strategy in cervical cancer radiation therapy. Acta Oncol. 2017;56(5):667-74. https://doi.org/10.1080/0284186X.2016.1275772.