Deep Reinforcement Learning for Personalized Dose Optimization in Oncology Treatment

Author: Maria Stevenson **Affiliation:** Department of Information Technology, Harvard University

Email: maria.stevenson@harvard.edu

Abstract

Personalized dosing in oncology offers the promise of maximizing therapeutic benefit while minimizing toxicity, yet clinical practice remains constrained by population-level guidelines and limited individualized decision support. Deep reinforcement learning (DRL) which combines representation learning with sequential decision-making under uncertainty provides a principled framework for learning individualized, time-varying dose policies from longitudinal patient data and simulated environments. This article develops a comprehensive, scholarly, and application-oriented treatment of DRL for personalized dose optimization in oncology. We synthesize theoretical foundations, model architectures, environment and reward design, safety and interpretability considerations, evaluation protocols, and translational pathways toward clinical deployment. We critically review opportunities and limitations, present methodological best practices, and propose a research and validation roadmap bridging preclinical simulation, retrospective evaluation, and prospective trials. Throughout, we ground discussion in established work on reinforcement learning and clinical decision support and highlight domain-specific challenges in oncology (heterogeneous tumor biology, delayed outcomes, sparse labels, and strong safety constraints). This manuscript is intended as a near-submission-ready review + methods article for researchers developing DRL-driven precision dosing systems in cancer care.

Keywords: deep reinforcement learning, personalized dosing, chemotherapy, precision oncology, sequential decision-making, safe RL, causal inference, simulation, translational AI.

1. Introduction

Precision medicine in oncology seeks to tailor diagnostic and therapeutic decisions to individual patient characteristics, tumor genomics, treatment history, and dynamic responses (e.g., tumor markers, imaging, toxicity trajectories). Dose optimization is a central component: for cytotoxic chemotherapies, targeted agents, and immunotherapies, the trade-off between efficacy and toxicity is both patient-specific and time-varying. Current dosing paradigms often rely on population-level metrics (e.g., body surface area, fixed schedules, or toxicity-driven reductions) and do not fully leverage longitudinal data collected during therapy.

Reinforcement learning (RL) provides a mathematical framework for sequential decision-making where an agent learns a policy to maximize cumulative reward in an environment characterized by state transitions and delayed outcomes. Deep RL (DRL) augments RL with deep function approximators (neural networks) to handle high-dimensional state spaces and complex dynamics (Mnih et al., 2015; Sutton & Barto, 2018). In healthcare, DRL has shown promise in sepsis

management, mechanical ventilation scheduling, and dynamic treatment regimes (Komorowski et al., 2018; Murphy, 2003). Applying DRL to oncology dose optimization allows learning of individualized dosing policies that adapt to evolving biomarkers, toxicity, tumor response, and comorbidities while explicitly optimizing long-horizon outcomes.

This article elaborates on DRL methods tailored for oncology dosing: problem formalization, state and reward design, model classes, offline and online learning strategies, safety and interpretability mechanisms, evaluation metrics, and pathways to clinical translation. We aim to provide enough methodological depth and rigor to support researchers and clinicians in designing reproducible, safe, and clinically relevant DRL systems.

2. Background and Related Work

2.1 Reinforcement learning fundamentals

Reinforcement learning formalizes the interaction between an agent and an environment as a (partially observable) Markov decision process (MDP) or POMDP when observations are noisy or incomplete. At each timestep t, the agent observes a state s_t, takes an action a_t, receives a reward r_t, and the environment transitions to s_{t+1}according to dynamics $P(s_{t+1} \mid s_t, a_t)$. The objective is to learn a policy $\pi(a \mid s)$maximizing expected discounted return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$(Sutton & Barto, 2018). In clinical dosing, states include clinical measurements, actions correspond to dose choices, and rewards reflect intermediate and final clinical outcomes.

2.2 Deep reinforcement learning

DRL leverages deep neural networks to approximate value functions, policies, or models of the environment (Mnih et al., 2015; Silver et al., 2016). Principal algorithmic families include value-based (e.g., DQN and its variants), policy-gradient (e.g., REINFORCE, PPO, A2C/A3C), and actor-critic methods (which combine value and policy estimators). Model-based RL attempts to learn dynamics and plan, often improving sample efficiency attractive in healthcare where real-world experimentation is costly.

2.3 RL in clinical settings

Applications of RL to healthcare have grown: Komorowski et al. (2018) demonstrated RL for sepsis treatment strategies using observational ICU data; other works apply RL to mechanical ventilation weaning, fluid management, and insulin dosing. Statistical work on dynamic treatment regimes (e.g., Q-learning, A-learning) provides important conceptual and methodological foundations for RL in sequential medical decision-making (Murphy, 2003; Shortreed et al., 2011). Oncology-specific applications are emerging, including adaptive radiotherapy scheduling and dose individualization in chemotherapy and targeted therapies, though the literature is still nascent.

2.4 Challenges unique to oncology dosing

Oncology poses specific challenges: outcomes (e.g., progression-free survival, overall survival) are long-horizon and censored; intermediate biomarkers (e.g., tumor size, circulating tumor DNA) provide informative but imperfect and sparse signals; toxicity events may impose safety constraints (dose-limiting toxicities, hospitalizations); patient heterogeneity across tumor histology and molecular subtypes induces complex interactions; and randomized experimentation for policy learning is ethically constrained. These features necessitate careful reward engineering, incorporation of uncertainty

and safety, strong causal reasoning, and reliance on retrospective data and validated simulators.

3. Problem Formulation: Dose Optimization as a Sequential Decision Task

3.1 State, action, and observation spaces

- **State (s_t)**: a rich, temporally-indexed representation of the patient including static covariates (age, sex, tumor subtype, genomic markers), dynamic covariates (lab values, vitals, tumor burden metrics, prior doses, toxicities), time-since-treatment, and latent variables inferred from prior history (e.g., estimated drug clearance). When full patient state is unobservable, model as a POMDP and use history-encoding methods (RNNs, Transformers) or belief-state estimators.
- **Action (a_t)**: discrete or continuous dosing decisions. Discrete actions might include dose levels (e.g., 0.5×, 1.0×, 1.5× standard dose), hold/titrate, or escalate/de-escalate. Continuous actions enable selecting precise milligram dosages but increase learning complexity and may require bounded action spaces.
- **Observation (o_t)**: clinically measurable surrogates at each visit (lab results, toxicity grades, imaging-derived metrics). Observations arrive irregularly; handle via time-aware models (time embeddings, decay-based imputations).

3.2 Transition dynamics and modeling

Cancer progression and pharmacodynamics/pharmacokinetics (PD/PK) determine transitions. Model-free RL ignores explicit dynamics, while model-based RL learns dynamics $P(s_{t+1} \mid s_t, a_t)$, enabling planning and counterfactual reasoning. Hybrid approaches combine mechanistic PK/PD models with data-driven components (physics-informed neural networks or Bayesian hierarchical models).

3.3 Reward design

A central design decision: reward must encode clinical priorities and long-term trade-offs. Candidate components:

- **Short-term clinical signals**: decreases in tumor markers, radiographic response, absence/reduction of grade ≥3 toxicities, preservation of quality-of-life indicators.
- **Long-term outcomes**: progression-free survival (PFS), overall survival (OS), time-to-progression (TTP). These are delayed and possibly censored.
- **Penalty terms**: toxicity penalties, hospitalization costs, severe adverse events, or large deviations from standard-of-care.

Construct composite reward $r_t = w_1 \cdot \Delta(\text{tumor burden}) + w_2 \cdot I(\text{toxicity}) + w_3 \cdot$ long-term proxy, where weights $w_i$reflect clinical priorities. Reward shaping can accelerate learning but must avoid introducing bias that favors short-term gains at long-term cost. Use domain experts to calibrate reward weights and validate with sensitivity analyses.

3.4 Constraints and safety

Clinical constraints are hard: avoid policies that risk severe toxicity. Formulate as constrained MDPs (CMDPs) with safety constraints on expected cumulative toxicity or use risk-sensitive objectives (CVaR optimization, worst-case regret). Safety layers (shielding, action filters), conservative policy improvement, and human-in-the-loop oversight are critical for deployment.

4. Data Sources: Retrospective Cohorts, Simulators, and Synthetic Data

4.1 Retrospective electronic health records (EHR), clinical trials, and registries

High-quality longitudinal datasets are necessary. Sources include institutional

oncology EHRs, multi-institutional registries, and clinical trial repositories. Challenges: missingness, variable measurement frequency, censoring, selection bias, and treatment assignment confounding. Preprocessing includes harmonization of variable definitions, time alignment, imputation strategies for missing data, and creation of time-series features.

4.2 Mechanistic and hybrid simulators

Simulators enable safe policy experimentation and counterfactual evaluation. Options:

- **Mechanistic PK/PD models**: well-established for many chemotherapies; simulate drug concentration-time profiles and toxicities.
- **Tumor growth models**: e.g., exponential, logistic, or agent-based tumor growth models calibrated to patient-level data.
- **Hybrid simulators**: combine mechanistic submodels (PK/PD, tumor kinetics) with stochastic patient-level variability (frailty terms) and data-driven residual models learned from observational data.

Calibration and validation against held-out clinical datasets is mandatory.

4.3 Synthetic data generation

Generative models (e.g., variational autoencoders, GANs, probabilistic graphical models) can augment training data, preserving privacy while providing diverse trajectories. Synthetic data must be validated to ensure distributional fidelity and not introduce artifacts.

5. Model Architectures and Algorithmic Choices

5.1 Offline (batch) vs online learning

- **Offline RL**: Learning from logged (historical) datasets without active experimentation. This is the most realistic initial setting for clinical applications. Offline RL must contend with distributional shift and limited support for optimal actions in historical data. Algorithms: Batch-constrained Q-learning (BCQ), Conservative Q-learning (CQL), and other conservative/regularized approaches mitigate overestimation and covariate shift.
- **Online RL**: Interactive learning in a live environment or simulator; permits exploration but requires safety constraints. In healthcare, online learning is primarily applicable in simulated environments or during carefully controlled clinical trials.

5.2 Value-based, policy-gradient, and actor-critic approaches

- **Value-based (DQN and variants)**: Suitable for discrete action spaces; approximate the Q-function using deep networks (Mnih et al., 2015). Use double-Q, dueling architectures, and prioritized replay to stabilize learning.
- **Policy-gradient / Actor-Critic (PPO, A3C, DDPG, SAC)**: Handle continuous action spaces and stochastic policies. Soft Actor-Critic (SAC) is sample-efficient and stable for continuous dosing. Proximal Policy Optimization (PPO) offers robustness in policy updates.
- **Distributional RL and uncertainty-aware methods**: Learn full return distributions (e.g., C51, QR-DQN) to quantify risk and inform conservative dose selection.

5.3 Representation learning for longitudinal clinical data

Time-series encoders like recurrent neural networks (GRU, LSTM), temporal convolutional networks, and Transformer-based models can encode patient histories. Incorporate time embeddings and masking to account for irregular sampling. Multi-modal fusion integrates structured EHR data, imaging features, and genomics.

5.4 Model-based RL and planning

Model-based methods learn a transition model and reward model to simulate counterfactual trajectories and perform planning (e.g., MPC, imagined rollouts). In oncology, integrating mechanistic PK/PD models as priors in model-based RL increases interpretability and sample efficiency.

5.5 Causal and counterfactual considerations

Causal inference methods (propensity score modeling, marginal structural models, structural nested models) are critical to address confounding in retrospective data and validate learned policies. Counterfactual regret estimation and off-policy policy evaluation (OPE) methods (importance sampling, weighted IS, doubly robust estimators) allow estimation of policy value from historical logs.

6. Training Protocols and Practical Considerations

6.1 Preprocessing and feature engineering

- Time discretization aligned with clinical decision frequency.
- Feature normalization, embedding of categorical variables.
- Missing data imputation: use clinically-informed methods (last observation carried forward where appropriate), multiple imputation, or model-based imputation using deep generative models.
- Construct history windows and summary statistics (e.g., slopes of tumor markers).

6.2 Offline RL: mitigating distributional shift

- Use conservative algorithms (CQL, BCQ) to avoid overconfident extrapolation to actions not present in the dataset.
- Regularize policies toward clinician behavior (behavior cloning augmentation) to ensure plausibility.
- Use uncertainty-aware policies: e.g., ensemble models, bootstrap ensembles, or Bayesian neural networks to quantify epistemic uncertainty.

6.3 Reward shaping and calibration

- Iteratively refine reward weights using clinician inputs and retrospective simulations.
- Ensure rewards are temporally consistent and avoid perverse incentives.
- Use multi-objective RL formulations or scalarization to balance efficacy and safety.

6.4 Hyperparameter tuning and validation

- Cross-validate via patient-based splits; avoid time leakage.
- Use multiple seeds and ensemble of models to ensure robustness.
- Evaluate sensitivity to reward weights, state representations, and action discretization.

7. Evaluation: Offline and Prospective Strategies

7.1 Offline evaluation metrics

- **Policy value estimates**: Off-policy evaluation (OPE) using weighted importance sampling, per-decision IS, or doubly robust methods.
- **Clinical surrogate outcomes**: simulated PFS, tumor shrinkage rates, cumulative toxicity incidence.
- **Safety metrics**: rate of high-toxicity actions, frequency of dosing outside standard-of-care bounds.
- **Robustness metrics**: performance under domain shift (e.g., across molecular subtypes).

7.2 Simulation-based validation

Use validated simulators to run counterfactual rollouts and compare policies under randomized seeds and patient heterogeneity. Simulators

permit stress-testing under rare but critical scenarios (e.g., organ failure).

7.3 Retrospective clinician-in-the-loop validation

Compare recommended actions to clinician decisions in held-out datasets; use domain experts to qualitatively assess policy plausibility. Conduct case studies where policies suggest divergent dosing to analyze rationale and potential benefits/risks.

7.4 Prospective evaluation and clinical trials

- **Phase 0 / Feasibility studies**: Evaluate safety and clinician acceptance in small cohorts with human oversight.
- **Randomized controlled trials (RCTs)**: Compare DRL-guided dosing versus standard-of-care or clinician-led dose selection for well-defined indications. Trial design must consider adaptive randomization, stopping rules, and ethical oversight.
- **Pragmatic trials and registries**: Post-deployment monitoring using registries and real-world evidence to evaluate long-term outcomes.

8. Safety, Interpretability, and Regulatory Considerations

8.1 Safety-by-design

- Constrain action spaces to clinically acceptable doses.
- Implement a safety filter: reject or flag any DRL-proposed dose that violates predefined clinical rules.
- Use conservative policy improvement techniques that only propose actions near the distribution of observed clinician actions until safety is established.

8.2 Interpretability and explanations

- Provide case-level explanations for dose recommendations: feature attributions (Integrated Gradients, SHAP), counterfactual examples (what-if analyses), and model-agnostic surrogates (decision trees approximating policy).
- Present uncertainty estimates and confidence intervals for proposed actions.

8.3 Human-in-the-loop and clinician workflows

Integrate recommendations as decision support rather than autonomous dosing: present the top k dose suggestions, rationale, key features driving the recommendation, and safety alerts. This respects clinician judgment and aids acceptance.

8.4 Regulatory and ethical landscape

- Regulatory agencies (FDA, EMA) require evidence of safety, effectiveness, and post-market surveillance for AI-enabled medical devices. Establish a clinical evidence plan and risk management documentation.
- Ethical concerns: fairness across patient subgroups, informed consent when using adaptive systems, transparency about algorithmic limitations.
- Data governance: ensure patient privacy, appropriate data use agreements, and model stewardship.

9. Case Study: Conceptual Example Pipeline

(This section outlines a full pipeline for a hypothetical targeted therapy dose optimization problem conceptual but detailed to guide implementation.)

1. **Problem scope**: Dose optimization for a tyrosine kinase inhibitor (TKI) in metastatic disease where therapeutic index varies across patients and adverse events are dose-limiting.
2. **Data collection**: Assemble multi-center retrospective cohort including dosing records, PK/PD measures, tumor response metrics, toxicity grades, prior therapies, and genomics.

3. **Simulator development**: Build a hybrid PK/PD + tumor kinetics simulator calibrated to cohort via hierarchical Bayesian inference. Validate simulator by reproducing distributions of PFS and cumulative toxicity in held-out patients.
4. **State representation**: Use a time-aware Transformer encoder over historical observations enriched with static covariates and posterior estimates from PK models.
5. **Action space**: Discrete actions mapping to multiples of the standard dose plus hold/reduce actions.
6. **Reward function**: Composite reward with immediate penalties for grade ≥3 toxicities and delayed rewards for tumor shrinkage and simulated PFS.
7. **Algorithm**: Offline DRL with Conservative Q-Learning (CQL) using ensemble Q-networks for uncertainty estimation and behavior-cloning regularization to limit extrapolation.
8. **Validation**:
   - OPE using doubly robust estimators on historical logs.
   - Simulator rollouts across patient strata.
   - Clinician review of top divergent cases.
9. **Deployment pathway**: Feasibility study with human oversight -> pilot randomized study comparing DRL-assisted dosing vs standard guideline-based dosing with safety endpoints.

This pipeline highlights the interplay among mechanistic modeling, conservative offline RL, and staged validation required for a credible translational project.

10. Limitations and Open Challenges

- **Causal confounding in observational data**: Incomplete adjustment can bias policy learning. Integration of causal inference methods is crucial but nontrivial.
- **Sparse and delayed outcomes**: Long-horizon endpoints and censoring complicate credit assignment; proxies and intermediate biomarkers help but may not fully capture long-term effects.
- **Limited action coverage in historical logs**: Offline learning is constrained by the diversity of actions in the dataset; conservative algorithms mitigate but do not eliminate this limitation.
- **Simulator fidelity**: Simulators are approximations; policies valid in silico may fail in vivo. Demand rigorous calibration and transparent uncertainty quantification.
- **Ethical and social considerations**: Algorithmic bias, patient autonomy, and clinician acceptance require ongoing stakeholder engagement.
- **Computational and data requirements**: High-quality multi-modal data and compute resources are needed; resource constraints may limit generalizability.

11. Recommendations and Best Practices

1. **Start with conservative offline RL**: Use CQL-style methods and behavior cloning regularization when learning from retrospective EHRs.
2. **Incorporate mechanistic knowledge**: Use PK/PD and tumor kinetics as priors or building blocks in model-based RL.
3. **Design clinically meaningful rewards**: Co-design reward functions with oncologists and iterate using sensitivity analysis.
4. **Prioritize safety and interpretability**: Implement action constraints, uncertainty quantification, and human-in-the-loop workflows.
5. **Use validated simulators**: For stress-testing and prospective trial design, ensure simulators reproduce real-world distributions.

6. **Adopt robust evaluation**: Combine OPE, simulator rollouts, clinician review, and ultimately prospective studies.
7. **Document model development and governance**: Maintain reproducible pipelines, versioning, and audit trails to satisfy regulatory and ethical requirements.

12. Discussion

DRL has the potential to transform oncology dosing by enabling individualized, adaptive, and temporally-aware policies that optimize long-term patient outcomes. The methodological toolbox offline DRL, model-based planning, uncertainty-aware policies, and causal inference offers a pathway to safe and effective decision support. However, substantial work remains in data curation, simulator fidelity, evaluation methodologies, and practical integration into clinician workflows. The translational path requires careful staging: rigorous offline validation, simulator-based safety testing, small feasibility studies, and well-designed randomized trials. Multidisciplinary collaboration among oncologists, pharmacologists, data scientists, and regulatory experts is indispensable.

13. Conclusion

This article provides a comprehensive, academically rigorous, and practically oriented blueprint for developing deep reinforcement learning systems for personalized dose optimization in oncology. By combining theoretical underpinnings, domain-specific modeling, and a conservative translational approach emphasizing safety, interpretability, and clinical validation, researchers can responsibly explore the potential of DRL in cancer treatment. Future work should focus on constructing high-fidelity simulators, integrating causal discovery methods, and executing prospective clinical evaluations that demonstrate improved patient-centered outcomes.

15. References

1. Fatunmbi, T. O. (2022). Leveraging robotics, artificial intelligence, and machine learning for enhanced disease diagnosis and treatment: Advanced integrative approaches for precision medicine. *World Journal of Advanced Engineering Technology and Sciences, 6*(2), 121–135. https://doi.org/10.30574/wjaets.2022.6.2.0057
2. Fatunmbi, T. O. (2021). Integrating AI, machine learning, and quantum computing for advanced diagnostic and therapeutic strategies in modern healthcare. International Journal of Engineering and Technology Research, 6 (1), 26–41. https://doi.org/10.34218/IJETR_06_01_002.
3. 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*, 1716–1720. *(peer-reviewed application of RL to critical-care dosing/therapy; serves as a benchmark for clinical RL methods.)*
4. Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.
5. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.
6. Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(2), 331–355.

7. Shortreed, S. M., Moodie, E. E. M., & Murphy, S. A. (2011). On dynamic treatment regimes and the assessment of their performance: With an application to alcoholism treatment trials. *Statistics in Medicine, 30*(13), 1516–1528.
8. (Representative) Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research* (application of deep policies; useful for representation discussions).
9. (Representative) Levine, S., et al. (2018). Learning hand-eye coordination for robotic grasping with deep learning and domain randomization. *IEEE/RSJ International Conference on Intelligent Robots and Systems*.
10. (Representative RL safety & offline methods) Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-Learning for Offline Reinforcement Learning. *(conference paper relevant algorithmic approach for offline healthcare RL).*
11. (Representative OPE & causal methods) Dudík, M., Langford, J., & Li, L. (2011). Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning*.