Reinforcement Learning Enhanced AI-AugETM for Adaptive Real-Time Dose Adjustment in Oncology Phase I Trials

Xuantianyi Feng
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
xuantianyi.feng877@buffalo.edu

Jack L. Taylor
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
jack403@ku.edu

Abstract

The acceleration of oncology drug development relies heavily on the efficiency and safety of Phase I dose escalation trials, which traditionally employ rule-based or model-guided designs to identify the maximum tolerated dose. Recent advances in artificial intelligence have introduced exposure-toxicity joint modeling frameworks such as AI-AugETM, which integrate pharmacokinetic and toxicity data to personalize dose recommendations. However, these systems operate under static assumptions and lack the capacity for real-time adaptation to evolving patient responses. This paper proposes a reinforcement learning enhanced extension of AI-AugETM that transforms the dose adjustment process into a continuous, adaptive decision-making framework capable of learning optimal dosing policies on the fly. We examine the architectural integration of reinforcement learning with the existing AI-AugETM infrastructure, focusing on state representation, reward design, and policy learning under high uncertainty. The system-level implications of deploying such an adaptive agent in regulated clinical environments are analyzed, including robustness to noisy data, trade-offs between exploration and patient safety, fairness across heterogeneous patient populations, and governance challenges related to interpretability and regulatory approval. Deployment considerations such as computational sustainability, integration with electronic health record systems, and real-time safety monitoring are discussed. By situating the technical innovation within a broader socio-technical context, this paper argues that reinforcement learning enhanced AI-AugETM offers a promising path toward truly personalized dose optimization, but its success hinges on carefully designed governance structures that balance algorithmic autonomy with clinician oversight. The framework also raises important questions about equity in algorithmic decision-making and the need for transparent, auditable models. We conclude by outlining a research agenda for future empirical validation and policy development.

Keywords

reinforcement learning, AI-AugETM, dose adjustment, Phase I clinical trials, oncology, adaptive systems, real-time decision support, algorithmic fairness.

1. Introduction

Phase I clinical trials in oncology represent the critical first step in translating preclinical discoveries into therapeutic interventions. These trials are designed to determine the safety and tolerability of new agents, typically by escalating doses across cohorts of patients until a predefined proportion experiences dose-limiting toxicity [2]. The conventional 3+3 design, while operationally simple, has been criticized for its inefficiency and inability to adapt to accumulating information during the trial [3]. Model-based designs such as the continual reassessment method [2], the Bayesian optimal interval design [3], and escalation with overdose control [4] have improved statistical efficiency, but they rely on parametric assumptions and are not inherently designed for real-time personalization. The introduction of artificial intelligence into this domain promises to address these limitations by leveraging high-dimensional data streams, including pharmacokinetic profiles, biomarkers, and dynamic toxicity grades [6].

The AI-AugETM framework [16] represents a significant step forward by jointly modeling exposure and toxicity in a manner that allows for patient-specific dose recommendations based on individual pharmacokinetic parameters. Nonetheless, the current implementation of AI-AugETM operates in a batch-update paradigm, where model parameters are re-estimated after each cohort, and the dose assignment rule is static until the next update. This approach does not fully exploit the temporal nature of patient responses, nor does it allow for continuous adjustment during the course of treatment. As clinical trials move toward adaptive and continuous monitoring paradigms [13], there is a growing need for decision support systems that can learn from ongoing interactions and refine their recommendations in real time without compromising safety.

Reinforcement learning (RL) provides a natural framework for sequential decision-making under uncertainty, where an agent interacts with an environment and learns a policy that maximizes cumulative reward [1]. In the context of dose adjustment, the RL agent observes patient states, selects dose levels, and receives feedback in the form of toxicity outcomes and, if available, efficacy signals. Recent work has demonstrated the feasibility of RL for personalized dosing in cancer therapy [15], yet most studies remain simulation-based and do not address the unique constraints of Phase I trials, such as small sample sizes, high safety stakes, and ethical oversight. Our contribution is to propose an architectural enhancement of AI-AugETM that embeds a reinforcement learning component to enable adaptive real-time dose adjustment. This paper focuses on the system-level implications rather than on a specific algorithm, analyzing the trade-offs, governance structures, and deployment challenges that arise when integrating RL into a regulated clinical workflow.

2. Background and Related Work

The landscape of dose-finding methodology has evolved from purely algorithmic designs to model-based Bayesian approaches that incorporate prior information and update posterior estimates after each patient or cohort [5]. The continual reassessment method [2] models the dose-toxicity relationship using a one-parameter logistic curve and updates the estimate of the maximum tolerated dose sequentially. The Bayesian optimal interval design [3] simplifies implementation by using predefined intervals for dose escalation, de-escalation, or stay decisions based on the observed toxicity rate. These methods have been extended to handle late-onset toxicities [8] and to jointly model efficacy and toxicity [9]. Despite their improvements, these designs operate at the cohort level and do not provide individual-level dose recommendations.

The emergence of artificial intelligence in clinical trial design has introduced new possibilities. Machine learning models have been used to predict toxicity from patient covariates, and exposure-toxicity modeling has been enriched by incorporating deep learning architectures [6]. The AI-AugETM framework [16] exemplifies this trend by jointly modeling the pharmacokinetic process and the binary toxicity response, enabling simulation of dose-exposure-toxicity relationships for individual patients. This allows the clinician to select a dose that achieves a target exposure window while controlling the predicted probability of toxicity. However, the framework relies on a fixed model trained on historical data or initial cohort data, and its recommendations do not adapt to unfolding toxicity events within the same patient or across patients in real time.

Reinforcement learning offers a solution to this static nature by framing dose adjustment as a Markov decision process. In healthcare, RL has been applied to sepsis management, fluid resuscitation, and personalized insulin dosing [7]. In oncology, RL has been explored for adaptive radiotherapy dosing and chemotherapy scheduling [15]. These applications highlight the potential of RL to learn complex policies from sequential data, but they also reveal significant challenges: the need for large amounts of interaction data, which is scarce in Phase I trials; the risk of unsafe exploration; and the difficulty of specifying a reward function that aligns with clinical objectives. Overfitting and distributional shift are additional concerns when the underlying patient population is heterogeneous [11]. The work by Gottesman et al. [7] provides guidelines for RL in healthcare, emphasizing off-policy evaluation, uncertainty quantification, and human oversight. Our proposal builds on these guidelines while specifically addressing the constraints of early-phase oncology trials.

3. The AI-AugETM Framework and Its Limitations

The AI-AugETM framework [16] integrates a pharmacokinetic model with a toxicity likelihood function to produce personalized dose predictions. The pharmacokinetic component estimates the time course of drug concentration in the body based on patient characteristics such as weight, renal function, and genetic factors. The toxicity component then maps the estimated exposure metrics, such as area under the curve or peak concentration, to the probability of experiencing a dose-limiting toxicity. By combining these two submodels, AI-AugETM can generate a risk profile for each candidate dose level and recommend the dose that maximizes the probability of staying within a predefined safety window. In a Phase I setting, this approach allows for more granular dose escalation than traditional cohort-based rules, potentially reducing the number of patients exposed to subtherapeutic or excessively toxic doses.

Nevertheless, the framework has several limitations that hinder its adaptability. First, the exposure-toxicity relationship is assumed to be stationary across patients and over time, yet biological factors such as drug accumulation, metabolic adaptation, and cumulative toxicity can alter this relationship as the trial progresses. Second, the model is typically trained on a limited dataset, and its predictions are not continuously updated based on incoming observations unless a full retraining step is performed. This batch-update process is computationally expensive and introduces a delay between data collection and model revision. Third, AI-AugETM does not incorporate the sequential nature of dose administration; it provides a single static recommendation for the next dose level, without considering that a patient may receive multiple doses over several cycles. In many oncology Phase I trials, patients are treated over multiple cycles, and the decision to escalate, maintain, or de-escalate

the dose should depend on the cumulative exposure and toxicity history. A static recommendation cannot capture these dynamics.

These limitations are not unique to AI-AugETM but are common among model-based dose-finding designs that do not employ a temporal learning component. By integrating reinforcement learning, we can transform AI-AugETM into a dynamic system that treats dose adjustment as a sequential decision process where the state includes not only baseline covariates but also the evolving history of exposures and toxicities. This enhancement addresses the key shortcomings and opens the door to real-time personalization.

4. Reinforcement Learning Enhancement Architecture

The proposed enhancement involves embedding a reinforcement learning agent on top of the existing AI-AugETM framework. The core idea is to replace the static dose rule with a policy learned through interaction. The RL agent observes a state that includes the patient's baseline covariates, the most recent exposure estimate from AI-AugETM's pharmacokinetic model, the cumulative toxicity score, and the time since the last dose. The action space consists of a discrete set of dose levels or, more flexibly, a continuous dose amount. After administering a dose, the environment produces a toxicity outcome, which may be binary or graded, as well as new pharmacokinetic data that update the exposure estimate. The reward signal is designed to balance safety and therapeutic benefit: a large negative reward for dose-limiting toxicities, a small positive reward for successfully completing a cycle without toxicity, and moderate rewards for achieving target exposure levels.

This architecture leverages the pretrained AI-AugETM model as a component of the state transition function. Specifically, the pharmacokinetic submodel can be used to simulate the expected exposure for a given dose and patient characteristics, and this simulated exposure becomes a deterministic element of the next state. However, the actual toxicity outcome is stochastic and depends on the true biological response. The RL agent must therefore learn to handle both model uncertainty and aleatoric uncertainty. To mitigate the risk of unsafe exploration, we propose incorporating a safety layer that overrides the agent's action if the predicted probability of toxicity from AI-AugETM exceeds a clinician-defined threshold. This safety layer is analogous to the notion of "guardrails" in autonomous systems and ensures that the agent does not recommend a dose that the existing static model deems unsafe [13].

The training of the RL agent poses significant challenges due to the small sample size of Phase I trials. Traditional RL relies on millions of interactions, which is infeasible in clinical settings. We therefore advocate for a simulation-based training strategy, where a digital twin of the trial environment is constructed using the AI-AugETM pharmacokinetic-toxicity model as a generative surrogate [10]. The agent is trained in this simulated environment to learn a policy, which is then fine-tuned with real data using off-policy methods. Uncertainty quantification techniques, such as Bayesian RL or ensemble methods, can provide confidence intervals on the agent's recommendations, enabling clinicians to make informed decisions about whether to follow the RL suggestion or revert to a conservative dose [7]. This hybrid approach respects the ethical imperative to prioritize patient safety while gradually introducing adaptivity.

5. Real-Time Dose Adjustment Mechanism

The integration of reinforcement learning enables a continuous feedback loop that updates dose recommendations as new data become available. In a typical Phase I trial, after each

patient receives a dose, toxicity and pharmacokinetic data are collected at multiple time points. The RL agent can process these data immediately after a predefined observation window, compute the updated state, and propose a dose for the next cycle or the next patient. This real-time adjustment stands in contrast to cohort-based designs where all patients in a cohort receive the same dose before any modification. The benefit is twofold: patients who experience early signs of toxicity can have their dose promptly reduced, and patients who tolerate the dose well can be escalated without waiting for the entire cohort to complete.

However, real-time adjustment introduces regulatory and practical complications. The trial protocol must specify the decision rule a priori, and an adaptive RL agent whose policy evolves during the trial may violate the principle of "operational characteristics" that regulators require to evaluate the design's type I error rates and safety properties [13]. To address this, the RL policy can be fixed at the time of trial registration, learned from extensive simulations, and then deployed as a deterministic decision rule. This preserves the adaptive nature without changing the policy during the trial, which is acceptable under current FDA guidance as long as the rule is prespecified [13]. A more ambitious approach would allow online learning with frequent interim analyses and oversight by a data safety monitoring board. We argue that for initial implementations, a fixed policy that has been validated in simulation is the most practical path to adoption.

Another critical aspect is handling missing data and irregular observation intervals. In real-world trials, pharmacokinetic samples may be missed, and toxicity evaluations may occur at non-standard times. The RL state representation must be robust to such irregularities, for example, by using a recurrent neural network to encode the time series of available observations [15]. The AI-AugETM exposure model can be used to impute missing exposure values, but the imputation uncertainty should be propagated into the policy evaluation. This uncertainty can be visualized for the clinician, who retains the ultimate decision authority. Thus, the system functions as a decision support tool rather than an autonomous agent.

6. System-Level Trade-Offs and Robustness

Deploying an RL-enhanced dose adjustment system involves navigating several trade-offs. The most fundamental is the exploration-exploitation dilemma: the agent must try different doses to learn about the dose-toxicity relationship, yet each exploration incurs a risk of exposing a patient to an unsafe dose. In Phase I trials, the acceptable risk level is extremely low, typically a 33% or lower rate of dose-limiting toxicity at the maximum tolerated dose. Therefore, the exploration must be heavily constrained. One approach is to use a "safe RL" framework that explicitly bounds the probability of selecting a dose with predicted toxicity above a threshold. This can be implemented by adding a risk constraint to the policy optimization objective, though such constraints can degrade learning efficiency [1].

A second trade-off is between personalization and statistical validity. Strong personalization may lead to a wide variety of doses being used across patients, complicating the estimation of the population dose-toxicity curve. This is a concern for regulators who rely on the aggregate data to determine the recommended Phase II dose. To mitigate this, the framework can produce a policy that personalizes only within a safety envelope while maintaining a systematic allocation to ensure that each dose level is adequately tested in a subset of patients. This is reminiscent of Bayesian adaptive design approaches that balance between individual and group objectives [5].

Robustness to model misspecification is another concern. The AI-AugETM model may be inaccurate for certain patient subgroups, leading to biased state estimates. If the RL agent relies heavily on these estimates, its policy may be suboptimal. To improve robustness, we propose using an ensemble of pharmacokinetic-toxicity models, each trained with different assumptions or on bootstrapped data, and feeding the aggregate state distribution to the RL agent. This ensemble approach also provides a measure of epistemic uncertainty that can be used to flag high-risk decisions for manual review [11].

Furthermore, the system must be robust to adversarial or non-stationary environments. For instance, if the patient population changes over time due to relaxed eligibility criteria, the dose-toxicity relationship may shift. The RL agent may need to detect such changes and adjust its policy. Continuous monitoring of the distributional drift can trigger a reassessment of the policy or a return to a conservative baseline [18]. Building such monitoring into the infrastructure is essential for long-term deployment.

7. Governance, Fairness, and Policy Considerations

The introduction of an algorithmic decision-maker in Phase I trials raises governance questions that extend beyond technical performance. Transparency and interpretability are paramount because clinicians, sponsors, and regulators must understand why a particular dose was recommended. The RL policy, often represented as a deep neural network, is inherently a black box. To address this, we advocate for the use of attention mechanisms or feature importance methods that can highlight which patient attributes contributed most to the recommendation. Additionally, the system should output confidence intervals and a counterfactual explanation showing what dose would have been recommended under alternative reasonable assumptions [17].

Fairness is a critical dimension in the context of oncology trials, where patient demographics, genetic ancestry, and comorbidities can influence drug exposure and toxicity. Biases encoded in the training data or the AI-AugETM model may lead to systematically different dose recommendations across populations, potentially exacerbating health disparities [11]. For example, if the pharmacokinetic model was built primarily on data from Caucasian patients, it may underestimate toxicity in Asian populations due to differences in drug metabolism. The RL agent, learning from such a model, could unduly escalate doses for underrepresented groups. To prevent this, the system must incorporate fairness constraints during policy learning, such as ensuring that the expected toxicity rate does not differ significantly across pre-specified demographic groups. Moreover, the trial governance structure should include a diverse ethics committee that reviews the algorithm's recommendations for any signs of implicit bias and has the authority to suspend the RL system if disparities emerge.

Regulatory policy must also evolve to accommodate RL-enhanced systems. Current FDA guidance on adaptive designs [13] does not explicitly address real-time machine learning algorithms. Regulators will demand rigorous pre-trial simulation studies, including sensitivity analyses under worst-case assumptions about model accuracy and patient heterogeneity. Post-trial, the algorithm's decisions should be audit-trailed and made available for independent review. The concept of "algorithmic accountability" in clinical trials requires that the manufacturer of the decision support system assume liability for failures, which may be a barrier to adoption. We recommend that early adopters collaborate with regulatory agencies to develop a qualification pathway for such systems, perhaps as a companion device to the investigational drug [17].

8. Deployment and Sustainability Challenges

Deploying an RL-enhanced AI-AugETM system in a real clinical trial setting involves significant infrastructural hurdles. The system must be integrated with the electronic health record and clinical data management systems to receive real-time data feeds of pharmacokinetic results and toxicity assessments. Data latency, missing values, and formatting inconsistencies must be handled gracefully by the RL agent, which may require a sophisticated data preprocessing pipeline that runs continuously. Computational sustainability is also a concern: training and inference for deep RL models can be resource-intensive, especially if the agent is updated online. Using smaller, more interpretable models such as linear function approximation in conjunction with a safety layer may be more feasible for low-resource settings [1].

The human-machine interaction design is equally important. Clinicians must trust the system's recommendations without ceding clinical judgment entirely. The interface should display the recommended dose, the predicted probability of toxicity, the current state, and a list of alternative doses with their associated risks. The system should also provide a "dose hold" or "reduce" option that the clinician can override. Studies in computer-supported cooperative work suggest that overly intrusive automation can lead to "alert fatigue" or distrust, so the system should provide recommendations only when they deviate from a clinician's expected action by a significant margin [12]. A gradual rollout, beginning with a silent mode where the system logs its recommendations but does not display them, can build confidence.

Sustainability also involves the long-term maintenance of the underlying AI-AugETM model and the RL policy. As new drugs and patient populations are introduced, the models may need to be updated. Continuous validation against accumulating real-world evidence is necessary to detect performance degradation. This raises the question of when an update triggers the need for a new regulatory submission. A pragmatic approach is to treat the entire system as a locked algorithm that remains static for the duration of a given trial, with updates reserved for the next trial. Over time, a library of validated RL policies for different drug classes and trial designs could be developed, resembling the concept of clinical practice guidelines but for algorithmic decision support.

9. Conclusion

The combination of reinforcement learning with the existing AI-AugETM framework offers a powerful approach to achieving adaptive real-time dose adjustment in oncology Phase I trials. By transforming static, cohort-based dose recommendations into a sequential decision-making process, the proposed system can personalize dosing to each patient's evolving exposure-toxicity profile, potentially accelerating the identification of safe and effective doses while reducing the number of patients exposed to subtherapeutic or unsafe levels. However, the path to practical deployment is fraught with technical, regulatory, and ethical challenges. We have analyzed the architectural integration, the necessity of safety constraints, the trade-offs between exploration and safety, the imperative of fairness and transparency, and the governance structures required to ensure responsible use. The system must be rigorously validated in simulation before clinical deployment, and its recommendations must remain subject to human oversight. Future work should focus on empirical evaluation of the RL-enhanced framework using real-world trial data, development of interpretability tools tailored to clinical decision-making, and engagement with regulatory bodies to establish clear guidelines for algorithmic decision support in early-phase trials. Only through a multi-

stakeholder effort can the promise of AI-driven adaptive dose adjustment be realized in a manner that is safe, equitable, and sustainable.

References

1. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
2. O'Quigley, J., Pepe, M., & Fisher, L. (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. Biometrics, 46(1), 33–48.
3. Liu, S., Yuan, Y., & Chi, Y. (2015). Bayesian optimal interval design for dose finding in phase I clinical trials. Journal of the American Statistical Association, 110(511), 1045–1056.
4. Babb, J., Rogatko, A., & Zacks, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. Statistics in Medicine, 17(10), 1103–1120.
5. Berry, D. A. (2006). Bayesian clinical trials. Nature Reviews Drug Discovery, 5(1), 27–36.
6. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
7. Gottesman, O., Johansson, F. D., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. Nature Medicine, 25(12), 1890–1899.
8. Cheung, Y. K., & Chappell, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics, 56(4), 1139–1144.
9. Thall, P. F., & Cook, J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. Biometrics, 60(3), 684–693.
10. Barricelli, B. R., Casiraghi, E., & Fogli, D. (2019). A survey on digital twin: Definitions, characteristics, applications, and design implications. IEEE Access, 7, 167653–167671.
11. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
12. Zhang, Y., Chen, X., & He, L. (2020). Real-time monitoring and adaptive decision support in clinical trials using machine learning. Journal of Biomedical Informatics, 108, 103500.
13. Food and Drug Administration. (2020). Adaptive designs for clinical trials of drugs and biologics: Guidance for industry. U.S. Department of Health and Human Services.
14. Liu, M., & Wang, Y. (2021). Model-based dose-finding designs for oncology phase I trials: A review. Pharmaceutical Statistics, 20(4), 709–724.
15. Chen, Z., & Shen, J. (2022). Reinforcement learning for personalized dosing in cancer therapy: A survey. IEEE Transactions on Neural Networks and Learning Systems, 33(11), 6183–6199.
16. Wang, Y. (2025, August). AI-AugETM: An AI-augmented exposure–toxicity joint modeling framework for personalized dose optimization in early-phase clinical trials. In 2025 19th International Conference on Complex Medical Engineering (CME) (pp. 182-186). IEEE.

17. Karger, D., & Loh, W. (2023). Ethical considerations in AI-driven clinical decision support. Nature Digital Medicine, 6, 112.

18. Gupta, R., & Raftery, A. E. (2024). Bayesian hierarchical models for adaptive dose escalation with real-time updates. Journal of the Royal Statistical Society: Series C, 73(2), 345–368.

19. Lee, J. J., & Chu, C. R. (2020). Bayesian dose-finding designs: A review and comparison. Clinical Trials, 17(5), 503–513.

20. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.