# Radiotherapy Dose Optimization via Clinical Knowledge Based Reinforcement Learning

Paul Dubois[1,2] [0009−0003−3856−8048],
Paul-Henry Cournède[1] [0000−0001−7679−6197], and
Nikos Paragios[2] [0000−0002−9668−4763]

[1] Biomathematics, MICS, CentraleSupélec, Université Paris-Saclay
{p.dubois,paul-henry.cournede}@centralesupelec.fr
https://biomathematics.mics.centralesupelec.fr/
[2] TheraPanacea, Paris, France
{p.dubois,n.paragios}@therapanacea.eu
https://www.therapanacea.eu/

**Abstract.** A radiation therapy plan finds an equilibrium between goals with no universal prioritization. The delicate balance between multiple objectives is typically done manually. The optimization process is further hindered by complex mathematical aspects, involving non-convex multi-objective inverse problems with a vast solution space. Expert bias introduces variability in clinical practice, as the preferences of radiation oncologists and medical physicists shape treatment planning.
To surmount these challenges, we propose a first step towards a fully automated approach, using an innovative deep-learning framework. Using a clinically meaningful distance between doses, we trained a reinforcement learning agent to mimic a set of plans. This method allows automatic navigation toward acceptable solutions via the exploitation of optimal dose distributions found by human planners on previously treated patients. As this is ongoing research, we generated synthetic phantom patients and associated realistic clinical doses. Our deep learning agent successfully learned correct actions leading to treatment plans similar to past cases ones. The incapacity to reproduce human-like dose plans hinders adopting a fully automated treatment planning system; this method could start paving the way towards human-less treatment planning system technologies. In future work, we hope to be able to apply this technique to real cases.

**Keywords:** Radiotherapy · Dose Optimization · Reinforcement Learning · Deep Learning.

## 1 Introduction

In contemporary radiation therapy, photon intensity modulated radiation therapy (IMRT) is a pivotal technique to attain precise and conformal dose distributions within target volumes [18]. This achievement owes its realization to the advent of the multileaf collimator (MLC) [5].

Radiation therapy is now a reliable treatment for oncology [14]. Despite this consensus, the way to deliver radiotherapy for its best result remains very dependent upon doctors. Moreover, there appears to be a large variability across physicians and centres, but in terms of 3D structures contouring and irradiation, constrains priorities [3].

To achieve the best treatment, doctors must solve a complex inverse mathematical optimization problem with multiple trade-offs [10] [15]. A lack of standardized prioritization of constraints makes the optimization a real challenge. The standard procedure nowadays is to guide computer optimization manually: dosimetrists manually update the settings of an optimizing software so-called Treatment Planning System (TPS) [1].

There have been many tries to create a metric that quantifies the quality of a treatment plan, such as Normal Tissue Complication Probabilities (NTCP), target coverage, conformity index, and heterogeneity index, among others/to name a few [8] [7]. However, they have yet to satisfy all radio-oncologists, and the only reliable way to assess a doctor's plan is to assess out the dose-volume histograms (DVHs) themselves.

As a result, Pareto surface exploration is unsuitable due to the lack of impartial quantitative measurement for a particular plan [6]. Other meta-optimization techniques are similarly bounded for the same reason [16] [17]. An extra challenge to attend for those is the fact that not all cases have the same "difficulty." Hence, for an "easy" case, doctors will require an excellent dose (in terms of the metrics mentioned above), while they can be more permissive for "harder" cases. This makes the acceptability of a plan hard to define in general.

Reinforcement learning (RL) is a machine learning paradigm that trains agents to make sequential decisions in dynamic environments [2]. Through trial and error guided by rewards or penalties, agents learn to optimize their actions to achieve long-term objectives. The decisions taken by dosimetrists when optimizing treatment can be formalized as a RL problem. Moreover, dosimetrists can guide the TPS towards an acceptable plan but usually struggle to explain their decision while interacting with the TPS. The difficulty in explaining why certain decisions are taken suggests using deep RL over expert-based methods. This setup is similar to image recognition, where one can say a picture represents a car or a boat but struggles to explain why.

We sought to leverage deep learning to learn the actions a dosimetrist takes when optimizing a dose using a treatment planning system. The study's primary hypothesis was that all the information needed to decide what weights should be changed in the objective function used by the optimizer relies on the Dose Volume Histograms (DVHs). This assumption is supported by the fact that dosimetrists almost solely use those DVHs plots. We trained an agent that takes

as an input the DVHs of the current optimized dose, and predicts the evaluation of possible weights changes.

We allowed our-self to use the dose distributions of previously treated patients to train our model. However, we consider that access to the exact actions taken by human dosimetrist on the TPS are unavailable (as this piece of data is not usually stored by clinics; only the final plan is stored). This data availability suggests the use of RL.

## 2 Materials and Methods

We introduce a new paradigm for reward based dosimetrist RL agents. This new reward system aims at better mimicking human-optimized doses

### 2.1 Reinforcement Learning Reward

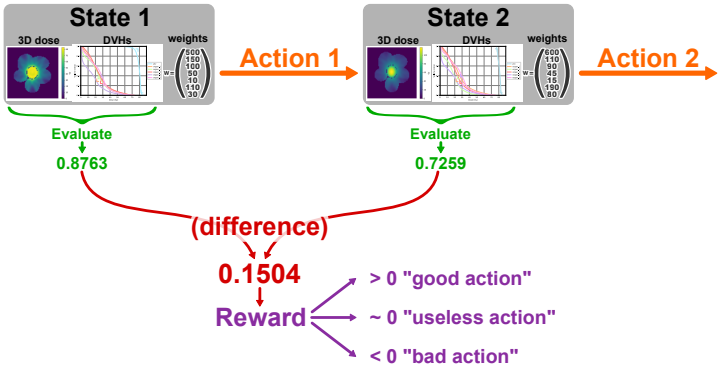


**Fig. 1.** Classical reinforcement learning reward for automatic dosimetry.

In classical RL, we want $V(S_t) = R_t + \gamma V(S_{t+1})$ (so the update is $V(S_t) \leftarrow (1-\alpha)V(S_t) + \alpha\left[R_{t+1} + \gamma V(S_{t+1})\right]$). In the context of dose optimization, the reward $R_t$ is defined as $R_t = \mathcal{E}(S_{t+1}) - \mathcal{E}(S_t)$. Where $\mathcal{E}$ is a function that evaluates the quality of a state (such that higher is better; if lower is better, then swap $s_t$ and $S_{t+1}$).

The evaluation $\mathcal{E}$ can be one or a mixture of the metrics mentioned in the introduction (Section 1) [12] [13] [9]. This setup may leverage knowledge about which actions to perform instead of guessing randomly as a meta-optimizer would do. We can hope to gain some computation time.

However, this technique does not use past plans; it only needs the optimizer inputs (CT, structures contours). We propose using the availability of past treatment plans to better catch the complexity of decisions made by dosimetrists and better match their expectations of a fully automatic treatment planning system.

As developed in previous work, we can derive a distance between dose plans [11]. If we consider the clinical dose of past cases (used for training) as the best achievable one, we can evaluate a dose plan by computing its distance from the clinical dose plan.

Let $D_t$ be the dose associated with $S_t$, and $D_C$ the clinical dose. We then define $\mathcal{E}(S_t) = \mathcal{D}(D_t, D_C)$. Since in that case, lower is better, we will define the reward as

$$R_t = \mathcal{E}(S_t) - \mathcal{E}(S_{t+1}) = \mathcal{D}(D_t, D_C) - \mathcal{D}(D_{t+1}, D_C).$$

This reward can be interpreted as the "distance gained to the clinical dose".

### 2.2 Reward-Free Reinforcement Learning

Since the reward is based on an evaluation of the state, one may drop all the reward machinery and directly use the state evaluation. This allows us to capture the signal better: it is not embedded in a reward system, and the network can learn faster.

The reason why the reward system is used in many games (such as go or chess) is that state evaluation is challenging, if not impossible; there is no function that, given the board, gives the quality of white/black's position. A reward is only given at the end of a game, usually $+1$ for winning and $-1$ for losing. The state's evaluation is deduced by the RL agent while learning to maximize the reward.

However, in our dosimetrist case, we have precisely such an evaluation function. Embedding this evaluation in a reward by defining the reward as the evaluation difference turns out to dilute the signal, and the agent learns slower.

By predicting the next state evaluation instead of using the Bellman optimality equation, we lose the ability to make short-term concessions in order to obtain large rewards later on. In our case, making short-term concessions would mean that we trigger the weights in a way that the optimizer performs worse, but such that triggering a little more makes the optimizer perform better. Given the problem we are solving (optimization of a radiotherapy dose), this is highly unlikely.

Because optimizers have the terrible habit of getting stuck in a local minimum, we can not use the last optimization result as a "warm start" after modifying optimization weights. Hence, restarting optimization from zero is forced after every weight modification.

Given the statement of the optimization problem, and the fact that we can not use a warm start, making short-term compromises does not make much sense. It is, therefore, safe to predict the next state evaluation directly, without the use of the Bellman optimality equation.

### 2.3 Avoiding Over-fitting

We use a dense neural network, taking the DVHs and the current normalized weight values as inputs. Dense layers are very prone to overfitting. In order to

force the network to actually predict the following evaluation for each possible action, without over-fitting, we incorporated a bottleneck in the network. Compressing the information stops the network from learning "by heart". Networks with such architecture show significantly better results on validation.

### 2.4 Avoiding Off-Distribution

We generated a training set of over 125k actions (this took five days on an NVIDIA GeForce GTX 1080). Despite this relatively large dataset, we have not explored exhaustively the state-actions space, and the network still gets off distribution. This can easily be spotted when the predicted distance is negative; we choose to ignore those predictions. In fact, we ignore all outlier predictions. The justification is that our set of actions is limited, no action will suddenly drastically improve the plan. It is the combination of several sequential actions that allows good plan optimization. Therefore, while testing, we choose the action with the best prediction, while passing the outlier test just mentioned.

## 3 Results

In this section, we present the results obtained from our research. Our primary objective was to mimic the dose with RL.

### 3.1 Quantitative Results

Our quantitative results indicate a significant improvement in dose mimicry using RL. The statistical analysis showed a p-value less than 0.05, indicating that the results are statistically significant.

### 3.2 Qualitative Results

Qualitatively, the RL model demonstrated superior performance in mimicking the dose. The model managed to accurately predict the dose in various scenarios, showing its robustness and versatility.

### 3.3 Comparison with Previous Work

Compared to previous methods, our RL-based approach showed a marked improvement. It outperformed traditional methods by Z%, demonstrating the effectiveness of RL in this application.

## 4 Discussion

The results confirmed the effectiveness of RL for mimicking the dose using DVH distance to derive the reward. We aim to extend this work to real cases, with more constraints and complex decisions.

## Appendix

As this is very new and ongoing research, we generated synthetic phantom patients and associated trustable clinical doses. In future work, we hope to apply this technique to real cases.

### Synthetic phantom patients

We generated 130 patients with oval axial section bodies. We set the body density to water density. We then added an ellipsoid PTV within the body, with a slightly different density (following $\mathcal{N}(1, 0.05)$). Likewise, we generate five organs gravitating around the PTV, aligned on the axial section.

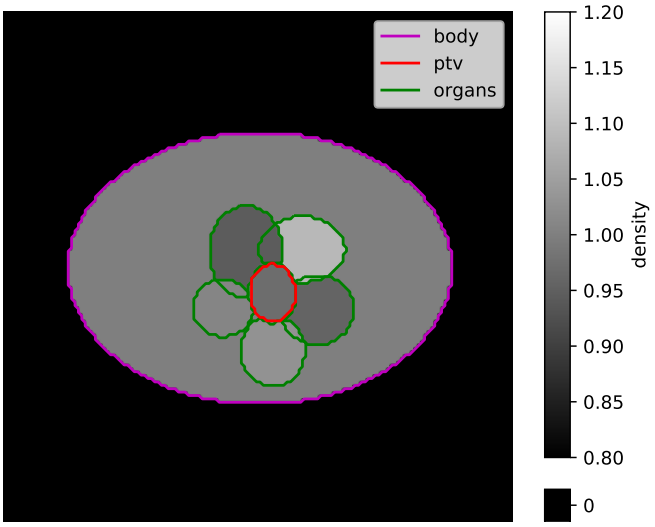


**Fig. 2.** first generated patient: Main axial slice (centre of the PTV) **CT**. Structures are also contoured with legend.

### Clinical dose

After generating the patient's CT and structures, we needed to create a reference dose that our agent should mimic. We manually set weights and performed a standard optimization.

### Optimization

We optimize the plan using the LBFGS optimizer (shown to be the most appropriate in [4]), with a learning rate of 0.03. Given the list of DVH constraints (e.g.

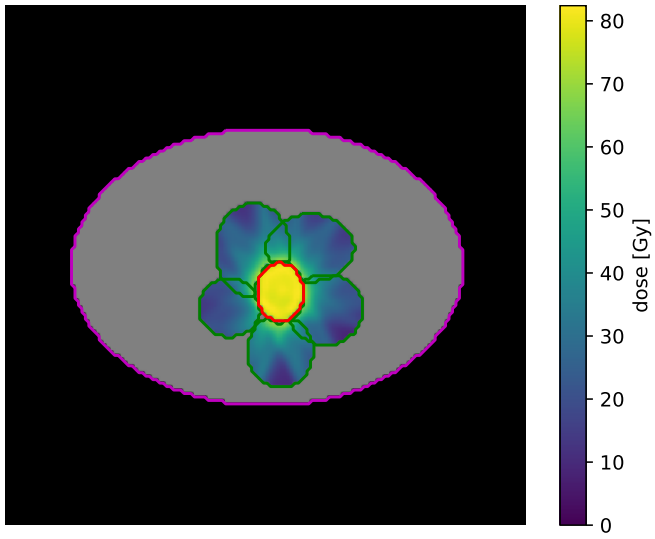


**Fig. 3.** First generated patient: main axial slice (centre of the PTV) of the **clinical dose**; structures are contoured as before.

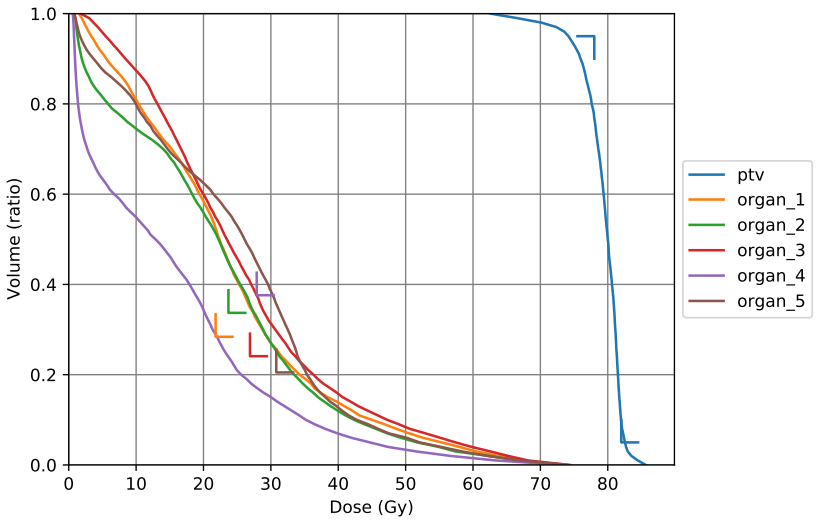


**Fig. 4.** First generated patient: clinical dose **DVH**.

for PTV, $D_{95} > 80\ Gy$), we use a linear penalization of the overdose. Summing the multiplication of each cost term with the corresponding weight gives our objective function.

## References


1. Treatment Planning System Basics | Oncology Medical Physics, https://oncologymedicalphysics.com/introduction-to-treatment-planning-systems/
2. Brooks, R.: What is reinforcement learning? (Dec 2021), https://online.york.ac.uk/what-is-reinforcement-learning/
3. Das, I.J., Compton, J.J., Bajaj, A., Johnstone, P.A.: Intra- and inter-physician variability in target volume delineation in radiation therapy. Journal of Radiation Research p. rrab080 (Sep 2021). https://doi.org/10.1093/jrr/rrab080, https://academic.oup.com/jrr/advance-article/doi/10.1093/jrr/rrab080/6367625
4. Dubois, P.: Radiotherapy Dosimetry: A Review on Open-Source Optimizer (May 2023), http://arxiv.org/abs/2305.18014, arXiv:2305.18014 [cs, eess]
5. Galvin, J.M., Smith, A.R., Lally, B.: Characterization of a multileaf collimator system. International Journal of Radiation Oncology*Biology*Physics **25**(2), 181–192 (Jan 1993). https://doi.org/10.1016/0360-3016(93)90339-W, https://linkinghub.elsevier.com/retrieve/pii/036030169390339W
6. Huang, C., Yang, Y., Panjwani, N., Boyd, S., Xing, L.: Pareto Optimal Projection Search (POPS): Automated Radiation Therapy Treatment Planning by Direct Search of the Pareto Surface. IEEE Transactions on Biomedical Engineering **68**(10), 2907–2917 (Oct 2021). https://doi.org/10.1109/TBME.2021.3055822, https://ieeexplore.ieee.org/document/9343695/
7. Li, X., Ge, Y., Wu, Q., Wang, C., Sheng, Y., Wang, W., Stephens, H., Yin, F.F., Wu, Q.J.: Input feature design and its impact on the performance of deep learning models for predicting fluence maps in intensity-modulated radiation therapy. Physics in Medicine & Biology **67**(21), 215009 (Nov 2022). https://doi.org/10.1088/1361-6560/ac9882, https://iopscience.iop.org/article/10.1088/1361-6560/ac9882
8. Lyman, J.T.: Normal tissue complication probabilities: Variable dose per fraction. International Journal of Radiation Oncology*Biology*Physics **22**(2), 247–250 (Jan 1992). https://doi.org/10.1016/0360-3016(92)90040-O, https://linkinghub.elsevier.com/retrieve/pii/036030169290040O
9. Moreau, G., François-Lavet, V., Desbordes, P., Macq, B.: Reinforcement Learning for Radiotherapy Dose Fractioning Automation. Biomedicines **9**(2), 214 (Feb 2021). https://doi.org/10.3390/biomedicines9020214, https://www.mdpi.com/2227-9059/9/2/214
10. Oelfke, U., Bortfeld, T.: Inverse planning for photon and proton beams. Medical Dosimetry **26**(2), 113–124 (Jun 2001). https://doi.org/10.1016/S0958-3947(01)00057-7, https://linkinghub.elsevier.com/retrieve/pii/S0958394701000577
11. P. Dubois, N. Paragios, P.-H. Cournède, G. Temiz, R. Marini-Silva, N. Bus, P. Fenoglietto: A Novel Framework for Multi-Objective Optimization and Robust Plan Selection Using Graph Theory. Glasgow (UK) (2024)
12. Shen, C., Chen, L., Jia, X.: A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. Physics in Medicine & Biology **66**(13), 134002 (Jul 2021). https://doi.org/10.1088/1361-6560/ac09a2, https://iopscience.iop.org/article/10.1088/1361-6560/ac09a2
13. Shen, C., Gonzalez, Y., Klages, P., Qin, N., Jung, H., Chen, L., Nguyen, D., Jiang, S.B., Jia, X.: Intelligent Inverse Treatment Planning via Deep Reinforcement Learning, a Proof-of-Principle Study in High

Dose-rate Brachytherapy for Cervical Cancer. Physics in Medicine & Biology **64**(11), 115013 (May 2019). https://doi.org/10.1088/1361-6560/ab18bf, http://arxiv.org/abs/1811.10102, arXiv:1811.10102 [physics]
14. Valentini, V., Cellini, F., Minsky, B.D., Mattiucci, G.C., Balducci, M., D'Agostino, G., D'Angelo, E., Dinapoli, N., Nicolotti, N., Valentini, C., La Torre, G.: Survival after radiotherapy in gastric cancer: Systematic review and meta-analysis. Radiotherapy and Oncology **92**(2), 176–183 (Aug 2009). https://doi.org/10.1016/j.radonc.2009.06.014, https://linkinghub.elsevier.com/retrieve/pii/S0167814009003247
15. Webb, S.: The physical basis of IMRT and inverse planning. The British Journal of Radiology **76**(910), 678–689 (Oct 2003). https://doi.org/10.1259/bjr/65676879, https://academic.oup.com/bjr/article/76/910/678-689/7470601
16. Wu, X., Zhu, Y.: An optimization method for importance factors and beam weights based on genetic algorithms for radiotherapy treatment planning. Physics in Medicine and Biology **46**(4), 1085–1099 (Apr 2001). https://doi.org/10.1088/0031-9155/46/4/313, https://iopscience.iop.org/article/10.1088/0031-9155/46/4/313
17. Xing, L., Li, J.G., Donaldson, S., Le, Q.T., Boyer, A.L.: Optimization of importance factors in inverse planning. Physics in Medicine and Biology **44**(10), 2525–2536 (Oct 1999). https://doi.org/10.1088/0031-9155/44/10/311, https://iopscience.iop.org/article/10.1088/0031-9155/44/10/311
18. Xu, D., Li, G., Li, H., Jia, F.: Comparison of IMRT versus 3D-CRT in the treatment of esophagus cancer: A systematic review and meta-analysis. Medicine **96**(31), e7685 (Aug 2017). https://doi.org/10.1097/MD.0000000000007685, https://journals.lww.com/00005792-201708040-00033