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]. However, 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 evaluate 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. The context-aware acceptability criteria make 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]. Agents learn to optimize their actions to achieve long-term objectives through trial and error guided by rewards or penalties. The decisions taken by dosimetrists when optimizing treatment can be formalized as an 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.

The study's primary hypothesis is 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). Our hypothesis is supported by the fact that dosimetrists almost solely use DVHs plots. In order to learn the actions of dosimetrists who use a TPS to optimize doses, we leverage deep learning. This is done by training an agent that takes the DVHs as the input of the current optimized dose, and predicts the evaluation of possible weights changes.

Typically, access to the exact actions taken by human dosimetrists on the TPS is unavailable (as clinics do not usually store this data; only the final plan is held). Therefore, we only use the dose distributions of previously treated patients to train our model. This partial availability of data 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 to improve how human-optimized doses are mimicked.

2.1 Reinforcement Learning Reward

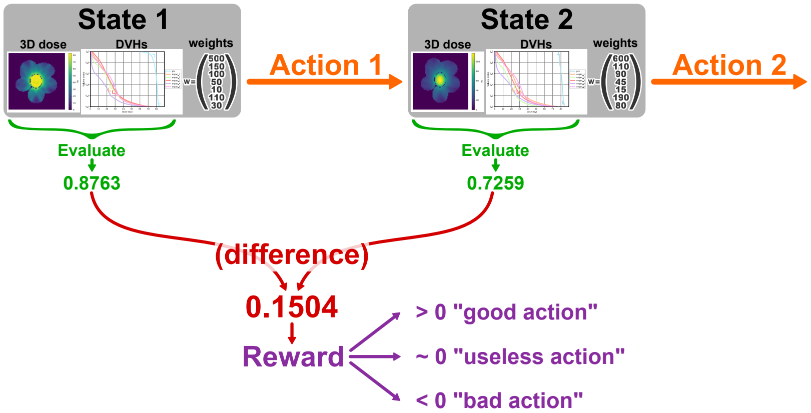

Fig. 1. Classical reinforcement learning reward for automatic dosimetry.

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. This could potentially 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 improve the detection of 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, $\mathcal{E}(S_t)$ should be minimized, 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 Architecture

We use a dense neural network, which inputs the DVHs and current normalized weight values. It outputs the $Q(s, a)$ value for each possible action a. Dense layers are very prone to overfitting. In order to force the network to actually predict the following evaluation for each possible action, without overfitting, we incorporated a bottleneck in the network (Ligure 2). Compressing the information stops the network from overfitting. Networks with such architecture show very little difference between training and validation sets (see Figure 2).

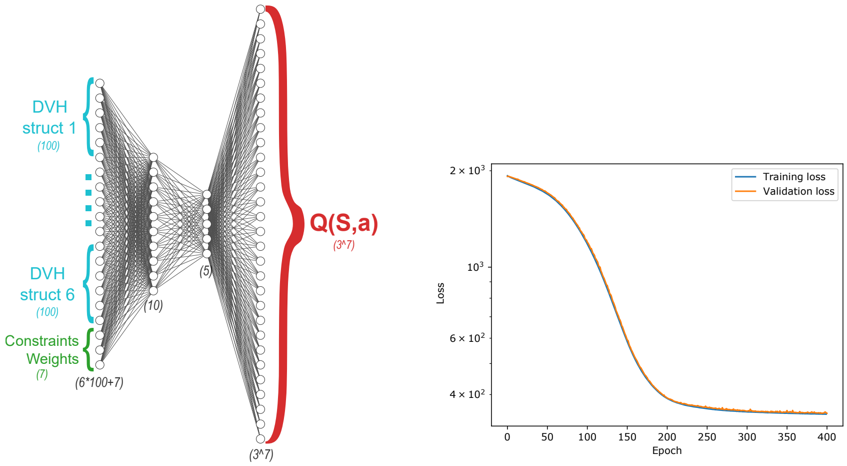

Fig. 2. Neural network architecture and loss evolution while training.

2.3 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, and in fact 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

Figure 3 shows how the distance between our RL agents performs over five steps on 30 test patients (unseen during the training). A lower distance is interpreted as an improved dose, since it is closer to the best dose, which is the clinical one.

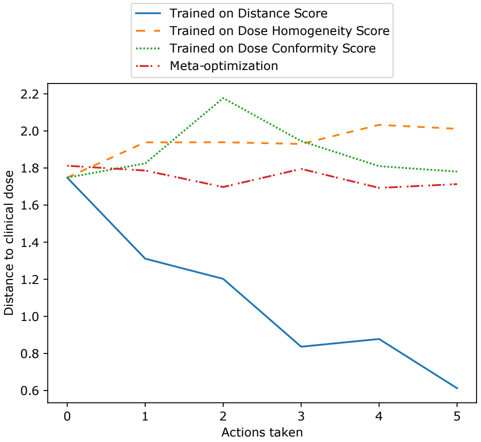

Fig. 3. Average distance between RL agent's dose and clinical dose.

3.1 Quantitative Results

The network converged on the training data, and validation showed minor overfitting. For testing, we generated 30 brand new cases that we again manually optimized. We then used the RL model to perform the optimization of these 30 unseen cases. On average, our model was able to reduce the dose distance with manually optimized dose by a factor of 2.5 (from 2.2 at iteration 0 to 0.91 at iteration 4), as shown in Table 3.1.

Agent \ Metric	Mean Final Distance*	Homogeneity Score†	Conformity Score†
RL Distance Score	**0.612**	1.871	0.406
RL Homogeneity Score	2.012	**4.387**	0.567
RL Conformity Score	1.781	4.232	0.451
Meta-optimization	1.279	4.117	**0.610**
Clinical doses	*0*	*1.541*	*0.580*

Table 1. Average performances of four algorithms tested on DVHs distance to clinical dose, dose homogeneity-based score, and conformity-based score.
**: distance is imporved performance through a lower score.*
†*: score is imporved performance through a higher score.*

3.2 Qualitative Results

Figure 4 shows the DVHs at each of the first four optimization steps on one of the test patients, unseen by the agent during the training. Our model drastically reduced the dose distance with manually optimized doses. Visual inspection of the DVHs plot shows that the dose optimized by the RL agent is very close to the clinical (manually fine-tuned) one.

4 Discussion

Our study demonstrates the potential of deep RL for automating radiotherapy treatment plan optimization. A key strength of our approach is its ability to learn from past treatment plans, capturing the complex decision-making processes of human dosimetrists. This data-driven approach avoids the limitations of predefined metrics, which may not fully capture the nuances of optimal treatment planning.

However, our study also has limitations. The agent's performance relies on the quality and quantity of available training data. Cases with limited historical data or complex anatomical features may require additional strategies. Moreover, while the agent achieves promising results regarding dose distance reduction, the dose is not guaranteed to be clinically acceptable. Although this study demonstrates the promise of our RL approach in a controlled setting, one final limitation to mention is that extending it to real-world radiotherapy planning would necessitates addressing additional complexities and constraints.

Several avenues exist for further research. Firstly, we plan to investigate strategies for incorporating additional information, such as patient characteristics and anatomical complexities, into the training process. Secondly, we aim to explore techniques for improving the interpretability of the agent's decision-making process, allowing for better understanding and potential clinical validation.

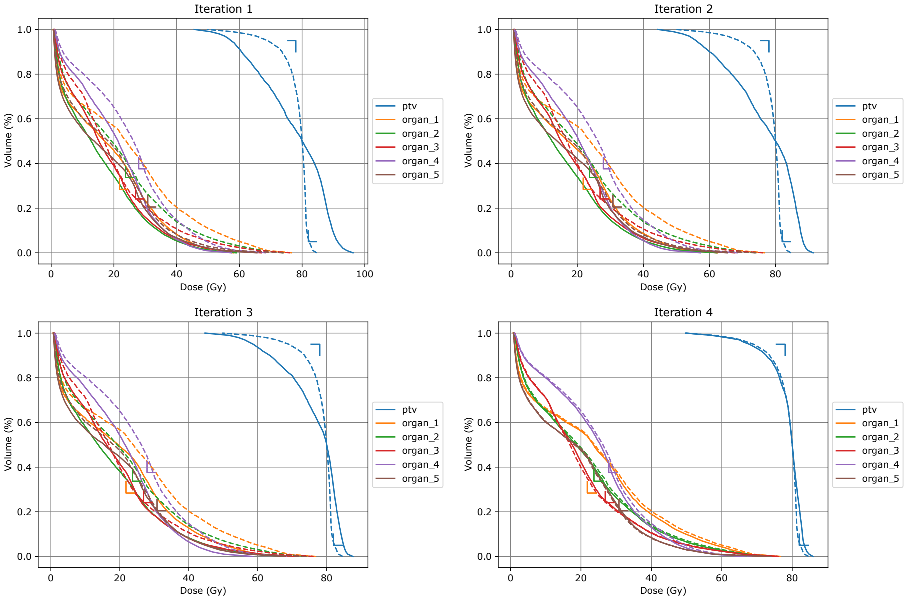

Fig. 4. RL Agent DVHs after each action taken on a test (unseen) patient. Solid lines are the agent's dose DVHs; dotted ones are the reference dose DVHs (manually fine-tuned).

5 Conclusion

Our approach differs from previous RL-based methods for radiotherapy planning in two key aspects. First, we avoid relying on pre-defined metrics for evaluation, which can be subjective, and limit the agent's ability to learn complex optimization strategies. Second, compared to traditional meta-optimization approaches, our method leverages past treatment data, potentially leading to more informed decision-making during the optimization process.

This study demonstrates deep RL's feasibility and potential benefits for automating radiotherapy treatment plan optimization. Our approach is capable of directly predicts state evaluations, and shows promise in achieving significant improvements in efficiency and, potentially, treatment outcomes. Further research is needed to address limitations, improve interpretability, and ensure safe clinical integration. This approach could revolutionize radiotherapy planning, leading to more standardized, efficient, and improved patient care.

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.

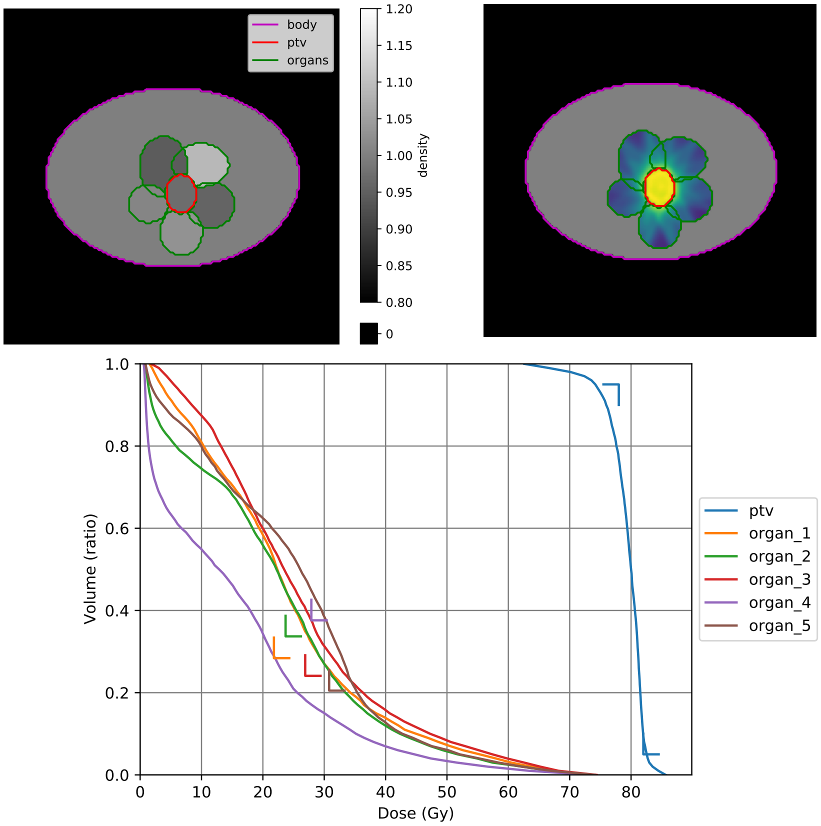

Fig. 5. Example of a (generated) patient:
Left: Main axial slice (center of the PTV) **CT**.
Right: Main axial slice (center of the PTV) of the **clinical dose**.
Bottom: Associated clinical dose **DVH**.

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. The dose prescription is a standard 80Gy on PTV, the same across all patients. We used a seven-beam IMRT irradiation technique on all the cohorts.

Optimization We optimize the plan using the LBFGS optimizer (shown to be the most appropriate in [4]). For each DVH constraint (e.g. for PTV, $D_{95} > 80\ Gy$), we used a linear penalization of the overdose.

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