Physics and Imaging in Radiation Oncology 28 (2023) 100496

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Evaluation of a clinically introduced deep learning model for radiotherapy treatment planning of breast cancer

Nienke Bakx [a], Maurice van der Sangen [a], Jacqueline Theuws [a], Johanna Bluemink [a], Coen Hurkmans [a,b,*]

[a] Department of Radiation Oncology, Catharina Hospital, Eindhoven, The Netherlands
[b] Faculties of Applied Physics and Electrical Engineering, Technical University Eindhoven, Eindhoven, The Netherlands

ARTICLE INFO

Keywords:
Breast cancer
Clinical use
Deep learning
Radiotherapy

ABSTRACT

Deep learning (DL) models are increasingly studied to automate the process of radiotherapy treatment planning. This study evaluates the clinical use of such a model for whole breast radiotherapy. Treatment plans were automatically generated, after which planners were allowed to manually adapt them. Plans were evaluated based on clinical goals and DVH parameters. Thirty-seven of 50plans did fulfill all clinical goals without adjustments. Thirteen of these 37 plans were still adjusted but did not improve mean heart or lung dose. These results leave room for improvement of both the DL model as well as education on clinically relevant adjustments.

1. Introduction

The process of radiotherapy treatment planning involves several manual and iterative steps, making it a time-consuming task. Besides, the outcome is prone to differences in experience of the planner [1]. To overcome these problems, in recent years, the number of studies to automate this process with the help of artificial intelligence (AI), and more specifically deep learning (DL), has increased [2,3]. The majority of these studies are of a retrospective nature, such as several studies for breast cancer [4–6], and only a limited number of DL models is actually implemented in clinical routine. However, a previous study regarding the clinical integration of DL treatment planning showed that a retrospective or simulated setting might not capture the real-world prospective setting of clinical decisions [7]. Therefore, the clinical use of such models should be monitored [8]. Recently, a DL model for dose prediction for left-sided whole breast radiotherapy was introduced, after thorough evaluation of the model in a pre-clinical study setting [9,10]. In this study, the real-word use of this model was monitored and evaluated to study the effect of using a DL model in a clinical setting.

2. Materials and methods

2.1. Patients

A DL dose prediction model, based on the 3D U-net architecture [11], was developed by RaySearch (RaySearch Medical Laboratories AB) and trained on an in-house collected dataset. This dataset was also previously used to train and evaluate a 2D U-net and a contextual atlas regression forest model [9,10]. However, since only the 3D U-net model is commercially available, this model was finally trained, commissioned and implemented in our clinical workflow in May 2022, using RayStation TPS. The dataset contained 105 left-sided node-negative breast cancer patients, treated with 15 fractions with a prescribed total dose of 40.1 Gy, in breath-hold position. A tangential IMRT was used, with a beam energy of either 6 or 10 MV, and up to 8 segments of at least 9 cm^2. Each beam contained at least one open segment with a high weight, together delivering approximately 200 MU, to promote robustness to swelling and breath hold position. More details on the dataset can be found in [9]. For evaluation, the clinical treatment plans of 50 patients treated after introduction of the DL model until December 2022 were included.

2.2. Treatment plan generation

The workflow to create DL plans consisted of several steps. First, the planner chose the appropriate beam energy, based on anatomy, and dorsal beam edges of the mediolateral and lateromedial beams were aligned. The beam angle was then automatically determined with the beam angle optimization method within RayStation, as previously described [12]. Next, the DL model predicted a voxel-wise dose

* Corresponding author at: Catharina Hospital, dept of Radiation Oncology, Eindhoven, The Netherlands.
E-mail address: coen.hurkmans@catharinaziekenhuis.nl (C. Hurkmans).

https://doi.org/10.1016/j.phro.2023.100496
Received 30 May 2023; Received in revised form 21 September 2023; Accepted 22 September 2023
Available online 27 September 2023

distribution, using a multichannel image volume as input, containing the binary masks of the PTV and OARs. However, since this prediction did not include any machine parameters and was not clinically applicable as such, dose mimicking was used to calculate a deliverable plan. The mimicking algorithm available in the TPS was used, of which technical details can be found in [9,10,13]. After mimicking, the leaves of the open tangential fields with a large contribution to the total dose were manually retracted from the skin surface by approximately 4 cm, and one last optimization run of 40 iterations for segment weight and shape was performed to generate the final plan, further referred to as DL plan. After this process, the planners were allowed to adjust the DL plan by their own discretion by adding extra objectives or changing weights of the objectives during plan optimization, regardless of whether the DL plan already fulfilled the clinical goals or not. These plans were further referred to as adjusted plans. If both the DL plan and adjusted plans did not fulfill the clinical goals, the planner created a treatment plan manually, following former clinical guidelines, referred to as manual plan.

2.3. Evaluation

For all patients, the final treatment plan was evaluated using predefined clinical goals of our institute, based on the Dutch national consensus criteria [14]. Furthermore, for all plans it was evaluated if the DL plan was directly used, adjusted, or if the final treatment plan was manually generated. If adjusted or manual plans were used, the original DL plans were generated again and compared to the final plans, based on several dose-volume histogram (DVH) parameters, such as mean dose to PTV, heart and lungs, and maximum dose to these regions, defined as the dose to 2% of the respective volume. To asses statistically significant differences, the Wilcoxon signed rank test was used.

3. Results

Fig. 1 summarizes the outcomes of the use of the DL model. For 37 patients (74%), the DL plans did fulfill all clinical goals without any adjustments. However, for 13 of these patients (35% of fulfilled plans), the treatment plan was still adjusted by the planner. Most adjustments were made to decrease high dose areas and resulted in a mean decrease of 0.3 Gy of the D2% of the PTV (range 0.0 – 0.7 Gy, $p < 0.05$). However, it did not affect the mean heart dose (MDH) or mean lung dose (MLD), since dose differences between DL and adjusted plans were within 0.02 Gy for both OARs. Thirteen treatment plans (26%) did not fulfill all clinical goals without intervention of the planner. In six cases (46% of failed plans), they could be easily adjusted to fulfill all clinical goals. In 1 case the MHD still exceeded the predefined clinical goal after adjustment, although it was improved after adjustments (3.0 Gy vs 2.5 Gy). Finally, for six patients a manual plan was created since the DL plan did not fulfil the clinical goals. The dose to 98% of PTV volume was insufficient for all these DL plans. Furthermore, the maximum dose to PTV was too high in four cases, too much of the external volume received a high dose in two cases and in one case the dose constraint of MHD was exceeded. For the latter case, manual planning still did not fulfill all clinical goals, but did improve PTV D2% (42.7 Gy vs 42.9 Gy) and MHD (2.2 Gy vs 2.4 Gy), compared to the DL plan. In addition, the beam angles were manually adjusted for two patients of this group, but re-planned DL plans with these angles still did not fulfill all clinical goals. In total, 45 of the patients were treated with a beam energy of 6 MV, resulting in five patients treated with 10 MV. For four of the patients treated with 10 MV a manual plan was created, while the DL plan of the fifth patient also needed manual adjustments before fulfilling all clinical goals. Median (range) PTV volumes for 6 and 10 MV patients were 854 cm^3 (316–2029) and 1738 cm^3 (1155–1939), respectively. Within the 6MV group, median volumes were 774 cm^3 (316–1332) for cases fulfilling all clinical goals, and 1155 cm^3 (736–2029) for the others.

The DVH parameters for DL and adjusted plans are shown in Table 1. A statistically significant difference was found for mean and maximum dose to PTV, whereas no difference was found for heart and lungs.

4. Discussion

This study evaluated the outcomes for the first 50 patients for which a DL model was clinically used to create treatment plans for left-sided whole breast radiotherapy. DL plans fulfilled predefined clinical goals in 74% of the cases without any adjustments, and in 86% of the cases with limited manual adjustments. Of the 20 plans that were manually adjusted, 13 plans (65%) already fulfilled clinical goals without these adjustments, without improving MHD and MLD. In general, manual adjustments only statistically significantly decreased mean and maximum dose to PTV. The 6 plans which were manually re-planned, contained 4 out of the 5 plans with a beam energy of 10 MV.

The difference in performance of the DL model for the two beam energies is noteworthy. When only considering 6 MV beam energy treatment plans, 82% of the DL plans fulfilled all clinical goals, which is 93% when also considering plans with small adjustments. The beam energy is chosen by the planner based on the patient anatomy, where 10 MV is chosen for a larger PTV volume. The DL plans that were manually re-planned, containing 4 out of the 5 10 MV plans, all lacked enough dose coverage to the PTV. When also considering the larger PTV volumes for the 6MV plans that did not fulfill all clinical goals, it could be stated that the DL model performs less for larger PTV volumes, and it should be investigated if separate mimick settings for this patient group would improve the dose coverage.

In our previous studies, two other DL models were tested both in a retrospective and pilot study, trained on the same dataset as the clinical DL model [9,10]. These were found to produce clinically acceptable

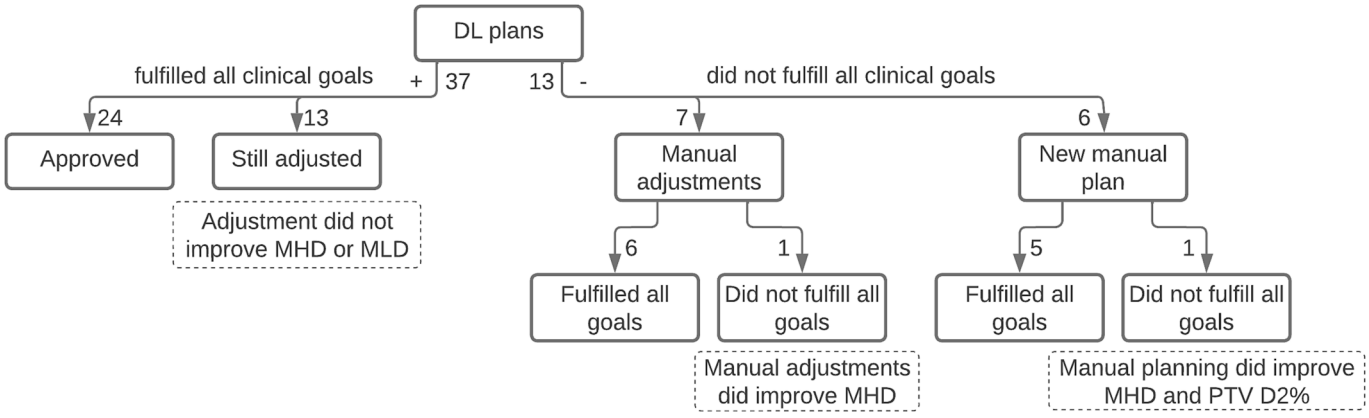

Fig. 1. Summary of the outcomes of the dl plans for 50 patients. mhd = mean heart dose, MLD = mean lung dose.

Table 1
Mean and maximum doses in Gy to PTV, heart and lungs for the DL and adjusted plans (mean ± standard deviation). An asterisk indicates a statistically significant difference ($p < 0.05$) between DL and adjusted plans.

	PTV				Heart		Lungs	
	Mean dose [Gy]		**Maximum dose [Gy]**		**Mean dose [Gy]**	**Maximum dose [Gy]**	**Mean dose [Gy]**	**Maximum dose [Gy]**
DL plan	40.4 ± 0.1	*	42.4 ± 0.3	*	1.1 ± 0.5	5.5 ± 6.6	2.3 ± 0.4	31.1 ± 3.1
Adjusted plan	40.3 ± 0.1		42.1 ± 0.4		1.1 ± 0.4	5.1 ± 4.9	2.3 ± 0.4	31.0 ± 3.2

plans in 90 to 95% of the cases, which is remarkably higher than the 74% found in this study. However, the clinical goals regarding mean dose to PTV were less strict, without a constraint on the maximum value, which leads to a higher acceptability rate. This difference stresses the importance to always take the current clinical context into account when evaluating such models.

No rejections of DL plans were reported in clinic, meaning all DL plans were deemed applicable when fulfilling all clinical goals. This result implies that the difference in clinical acceptability of the treatment plans between the pilot study and current clinical use is only caused by the above mentioned difference in clinical goals, in contrast to the study of McIntosh *et al.*, where the difference in acceptability is probably caused by a difference in perception of physicians towards the use of AI [7]. Conducting a pilot study and aligning quality standards for DL plans can therefore be regarded as requirements before successful clinical implementation of such models. To our knowledge, no other studies are monitoring DL models for planning in clinical practice to compare to, although Esposito *et al.* state that their approach is currently under clinical implementation [15].

This study shows that several aspects can be improved to further optimize the workflow and actually reduce the time needed for the whole treatment planning process. First of all, the outcome of the model could be further optimized by either improving the predicted or the mimicked dose. The outcomes of both steps can be steered by settings within the RaySearch algorithm, which were predefined during commissioning. Examples of settings are a minimum/maximum goal for ROIs after prediction, and minimum/maximum reference dose objectives, aiming to keep the dose at least/most at the predicted dose's levels during mimicking. Different mimick settings for different plan characteristics, such as beam energies as suggested before, could be tested. The predicted dose could also be improved by further training of the model, for example by including more data. Eventually, an improved predicted dose better reflects a clinically deliverable treatment plan, making it less depending on dose mimicking. It was also observed that 35% of the fulfilled plans were still manually adjusted, although it did not improve MHD or MLD and thus had no clinical relevance. These adjustments were mainly made to decrease high dose regions in the PTV after visual inspection, while the PTV D2% already did not exceed the clinical goal. This preference of tweaking of the dose between 95 and 107% of the prescribed dose (mean dose within 1% of prescribed dose) was previously already shown to be strongly observer dependent. Therefore, it is important to educate planners on the outcomes of such evaluation, to make them aware of the effect of adjustments and optimize the workflow by minimizing these adjustments.

In conclusion, the DL model successfully created a treatment plan in 74% of the cases without manual intervention, or in 86% of the cases when small manual adjustments are considered. Clinical results did not differ much from the pilot study. Improvement of training and configuration of the DL model is still possible, but discussion and education on clinically relevant adjustments is also of high importance.

CRediT authorship contribution statement

Nienke Bakx: Conceptualization, Software, Validation, Methodology, Writing - original draft. **Maurice van der Sangen:** Methodology, Writing - review & editing. **Jacqueline Theuws:** Methodology, Writing - review & editing**Johanna Bluemink:** Software, Validation, Methodology, Writing - review & editing. **Coen Hurkmans:** Conceptualization, Software, Validation, Supervision, Funding acquisition, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nienke Bakx received funding from RaySearch Laboratories AB. RaySearch Laboratories AB had no influence on the design of the study but aided with the practical implementation and by offering advice on practical issues.

Acknowledgements

We would like to thank Fredrik Löfman and the ML engineers from the machine learning department of RaySearch Laboratories AB for their contribution by assisting with the implementation and commissioning of the DL model. A research grant from Raysearch Laboratories AB is also acknowledged.

References

[1] Wang J, Hu W, Yang Z, Chen X, Wu Z, Yu X, et al. Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients? Radiat Oncol 2017;12:85. https://doi.org/10.1186/s13014-017-0822-z.
[2] Wang C, Zhu X, Hong JC, Zheng D. Artificial Intelligence in Radiotherapy Treatment Planning: Present and Future. Technol Cancer Res Treat 2019;18:1–11. https://doi.org/10.1177/1533033819873922.
[3] Wang M, Zhang Q, Lam S, Cai J, Yang R. A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning. Front Oncol 2020;10:580919. https://doi.org/10.3389/fonc.2020.580919.
[4] Ahn SH, Kim ES, Kim C, Cheon W, Kim M, Lee SB, et al. Deep learning method for prediction of patient-specific dose distribution in breast cancer. Radiat Oncol 2021; 16:154. https://doi.org/10.1186/s13014-021-01864-9.
[5] Hedden N, Xu H. Radiation therapy dose prediction for left-sided breast cancers using two-dimensional and three-dimensional deep learning models. Phys Med 2021;83:101–7. https://doi.org/10.1016/j.ejmp.2021.02.021.
[6] Bai X, Zhang J, Wang B, Wang S, Xiang Y, Hou Q. Sharp loss: a new loss function for radiotherapy dose prediction based on fully convolutional networks. Biomed Eng Online 2021;20:101. https://doi.org/10.1186/s12938-021-00937-w.
[7] McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27:999–1005. https://doi.org/10.1038/s41591-021-01359-w.
[8] Barragán-Montero A, Bibal A, Dastarac MH, Draguet C, Valdés G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model dependency. Phys Med Biol 2023;67:11TR01.. https://doi.org/10.1088/1361-6560/ac678a.
[9] Bakx N, Bluemink H, Hagelaar E, Van Der SM, Theuws J, Hurkmans C. Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer. Phys Imaging Radiat Oncol 2021;17:65–70. https://doi.org/10.1016/j.phro.2021.01.006.
[10] Kneepkens E, Bakx N, van der Sangen M, Theuws J, van der Toorn PP, Rijkaart D, et al. Clinical evaluation of two AI models for automated breast cancer plan generation. Radiat Oncol 2022;17:25. https://doi.org/10.1186/s13014-022-01993-9.
[11] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Med Image Comput Comput Assist Interv 2016;9901:424–32. https://doi.org/10.1007/978-3-319-46723-8_49.
[12] Bakx N, Bluemink H, Hagelaar E, van der Leer J, van der Sangen M, Theuws J, et al. Reduction of heart and lung normal tissue complication probability using

automatic beam angle optimization and more generic optimization objectives for breast radiotherapy. Phys Imaging Radiat Oncol 2021;18:48–50. https://doi.org/10.1016/j.phro.2021.04.002.

[13] Borderias-Villarroel E, Huet Dastarac M, Barragán-Montero AM, Helander R, Holmstrom M, Geets X, et al. Machine learning-based automatic proton therapy planning: Impact of post-processing and dose-mimicking in plan robustness. Med Phys 2023;50:4480–90. https://doi.org/10.1002/mp.16408.

[14] Hurkmans C, Duisters C, Peters-Verhoeven M, Boersma L, Verhoeven K, Bijker N, et al. Harmonization of breast cancer radiotherapy treatment planning in the Netherlands. Tech Innov Patient Support Radiat Oncol 2021;19:26–32. https://doi.org/10.1016/j.tipsro.2021.06.004.

[15] Esposito PG, Castriconi R, Mangili P, Broggi S, Fodor A, Pasetti M, et al. Knowledge-based automatic plan optimization for left-sided whole breast tomotherapy. Phys Imaging Radiat Oncol 2022;23:54–9. https://doi.org/10.1016/j.phro.2022.06.009.