advances
in radiation oncology

www.advancesradonc.org

Scientific Article

Machine Learning for Predicting Clinician Evaluation of Treatment Plans for Left-Sided Whole Breast Radiation Therapy

Christian Fiandra, PhD,[a,*] Federica Cattani, MSc,[b] Maria Cristina Leonardi, MD,[c] Stefania Comi, MSc,[b] Stefania Zara, MSc,[d] Linda Rossi, PhD,[e] Barbara Alicja Jereczek-Fossa, MD, PhD,[c,f] Piero Fariselli, PhD,[g] Umberto Ricardi, MD,[a] and Ben Heijmen, PhD[e]

[a]*Department of Oncology, University of Turin, Turin, Italy;* [b]*Unit of Medical Physics, IEO European Institute of Oncology IRCCS, Milan, Italy;* [c]*Division of Radiation Oncology, IEO European Institute of Oncology IRCCS, Milan, Italy;* [d]*R&D Department, Tecnologie Avanzate, Turin, Italy;* [e]*Department of Radiation Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands;* [f]*Department of Oncology and Hemato-oncology, University of Milan, Milan, Italy; and* [g]*Department of Medical Sciences, University of Torino, Turin, Italy*

Received 19 January 2023; accepted 15 March 2023

Purpose: The objective of this work was to investigate the ability of machine learning models to use treatment plan dosimetry for prediction of clinician approval of treatment plans (no further planning needed) for left-sided whole breast radiation therapy with boost.

Methods and Materials: Investigated plans were generated to deliver a dose of 40.05 Gy to the whole breast in 15 fractions over 3 weeks, with the tumor bed simultaneously boosted to 48 Gy. In addition to the manually generated clinical plan of each of the 120 patients from a single institution, an automatically generated plan was included for each patient to enhance the number of study plans to 240. In random order, the treating clinician retrospectively scored all 240 plans as (1) approved without further planning to seek improvement or (2) further planning needed, while being blind for type of plan generation (manual or automated). In total, 2 × 5 classifiers were trained and evaluated for ability to correctly predict the clinician's plan evaluations: random forest (RF) and constrained logistic regression (LR) classifiers, each trained for 5 different sets of dosimetric plan parameters (feature sets [FS]). Importances of included features for predictions were investigated to better understand clinicians' choices.

Results: Although all 240 plans were in principle clinically acceptable for the clinician, only for 71.5% was no further planning required. For the most extensive FS, accuracy, area under the receiver operating characteristic curve, and Cohen's κ for generated RF/LR models for prediction of approval without further planning were 87.2 ± 2.0/86.7 ± 2.2, 0.80 ± 0.03/0.86 ± 0.02, and 0.63 ± 0.05/0.69 ± 0.04, respectively. In contrast to LR, RF performance was independent of the applied FS. For both RF and LR, whole breast excluding boost PTV ($\text{PTV}_{40.05\text{Gy}}$) was the most important structure for predictions, with importance factors of 44.6% and 43%, respectively, dose recieved by 95% volume of $\text{PTV}_{40.05}$ ($D_{95\%}$) as the most important parameter in most cases.

Conclusions: The investigated use of machine learning to predict clinician approval of treatment plans is highly promising. Including nondosimetric parameters could further increase classifiers' performances. The tool could become useful for aiding treatment planners in generating plans with a high probability of being directly approved by the treating clinician.

Sources of support: This work had no specific funding.

Disclosures: Dr Zara is an employee of the company Tecnologie Avanzate TA Srl, which distributes the software RayStation in Italy; this company supports the group in terms of collection of data. No other disclosures were reported.

Research data are stored in an institutional repository and will be shared upon request to the corresponding author.

*Corresponding author: Christian Fiandra, PhD; E-mail: christian.fiandra@unito.it

https://doi.org/10.1016/j.adro.2023.101228

Introduction

Radiation therapy (RT) treatment planning for breast cancer focuses on reducing radiation exposure to healthy tissues (whole heart, left anterior descending coronary artery [LAD], lungs, and contralateral breast [CB]), while ensuring an adequate target coverage. Two phase 3 studies have shown significant toxicity reductions with intensity modulated RT (IMRT) compared with 3-dimensional (3D) conformal RT.[1,2] Apart from regular C-arm linear accelerators, static beam IMRT for patients with breast cancer can also be delivered with TomoDirect, an IMRT modality delivered with TomoTherapy (Accuray, Madison, WI).[3-6]

In a standard clinical practice, treatment plans are generated by planners and presented to the treating clinician for approval. Often, the final approved plan is the product of an iterative procedure in which an initial plan is stepwise enhanced to best satisfy the clinician's requirements. If on the one hand this can be a process that can avoid human errors,[7] it is also time-consuming and workload intensive.

Automated planning has been proposed to enhance plan quality and reduce workload.[8,9,28] However, several studies with blinded plan comparisons have shown that clinicians do not always prefer the automatically generated plan.[10-12] Recently, Cagni et al[13] systematically investigated differences in plan scoring among planners and treating clinicians in a single department. Large differences in plan quality assessments were observed.

In this study, we have investigated the ability of random forest (RF) or constraint logistic regression (LR) classifiers to use treatment plan dosimetry for correct prediction of clinicians' plan evaluations for left-sided whole-breast RT (WBRT) with boost as (1) approved without further planning to seek improvement or (2) further planning needed. The basis of the study was treatment plans for previously treated patients. To enhance the statistical power of the study, for each patient the manually generated clinical plan and an automatically generated plan were included. For study purposes, the involved clinician retrospectively labeled in random order all clinical and automatically generated study plans as (1) approved without further planning or (2) further planning needed, while being blinded for type of applied plan generation (manual or automated).

For both RF and LR, 5 different dosimetric feature sets (FS) were investigated (2 × 5 investigated classifiers in total) to assess dependence of prediction quality on selected plan parameters. Machine learning (ML) predictions for plans that were labeled "approved without further planning" were considered correct in case of a predicted probability P (approved without further planning) $>.5$.

For each of the investigated 2 × 5 classifiers, nested cross-validation was used to establish both hyperparameters and assess model performance, using the same data set.[14] Importance of included features for predictions was investigated to better understand in clinicians' plan evaluations.

To the best of our knowledge, this study is the first attempt of using ML with dosimetric plan parameters as input to predict clinicians' plan evaluations. In a hypothesized future clinical application, a planner could then first assess the probability that the clinician would consider a generated plan approved. If this probability is low, the planner could then try to further improve the plan before presenting it to the clinician, thereby minimizing the time used by clinicians for plan evaluations.

Methods and Materials

Patient selection and treatment planning

A total of 120 patients receiving adjuvant left-sided WBRT after breast-conserving surgery at the European Institute of Oncology (IEO) Institute between 2019 and 2020 were randomly selected from the institutional database. The study approved by the Ethical Committee of the IEO Institute (identification number UID2433). Institute (identification number UID2433). RT was delivered with TomoDirect in a TomoTherapy Hi-Art System (Accuray, Sunnyvale, CA).

Clinical plans were manually generated with the VOLO treatment planning system (version 2.1.6; Accuray, Sunnyvale, CA), applying a jaw width of 2.5 cm, a pitch of 0.25, and modulation factors of 1.8 to 2.0 to keep the delivery time within the range of 10 to 15 minutes. Breast and tumor bed were contoured based on European Society for Therapeutic Radiation and Oncology guidelines for early breast cancer.[15] Isotropic 5-mm expansions were added to create the corresponding planning target volumes (PTVs). Organs at risk (OARs) included left and right lung, CB, heart, and LAD.[16] In line with the Radiation Therapy Oncology Group 1005 study protocol,[17] 40.05 Gy was delivered to the whole breast in 15 fractions over 3 weeks with a simultaneously integrated boost to the tumor bed that resulted in a total dose of 48 Gy. Dose objectives mainly followed those used in the previously mentioned protocol (Table 1).

Table 1 Dose-volume histogram constraints for clinical planning and recommended and maximum acceptable values for all considered targets and organs at risk

Organ at risk	Ideal	Acceptable
Heart	V_{16Gy} <5%	V_{20Gy} <5%
	V_{8Gy} <30%	V_{8Gy} <35%
	D_{mean} <32 Gy	D_{mean} <4 Gy
Left anterior descending coronary artery	D_{mean} <25 Gy	
	$D_{1\%}$ <45 Gy	
Left lung	V_{16Gy} <15%	V_{16Gy} <20%
	V_{8Gy} <35%	V_{8Gy} <40%
	V_{4Gy} <50%	V_{4Gy} <55%
	D_{mean}	
Right lung	V_{4Gy} <10%	V_{4Gy} <15%
Contralateral breast	$D_{0,03cc}$ <2.4 Gy	$D_{0,03cc}$ <3.84 Gy
	$D_{5\%}$ <1.44 Gy	$D_{5\%}$ <2.40 Gy
	D_{mean} <4 Gy	D_{mean} <5 Gy
$PTV_{40.05Gy}$ (whole breast − boost volume)	$D_{95\%}$ >38 Gy	$V_{90\%}$ >36 Gy
	$D_{50\%}$	
	$D_{30\%}$	
	$D_{0,03cc}$ ≤46 Gy	$D_{0,03cc}$ ≤48 Gy
	CI	
	HI	
	$D_{95\%}$ ≥45.6 Gy	$D_{90\%}$ ≥43.2 Gy
$PTV_{48.0Gy}$ (boost volume)	$D_{5\%}$ ≤52.8 Gy	$D_{10\%}$ ≤52.8 Gy
	$D_{0,03cc}$ ≤55.2 Gy	$D_{0,03cc}$ ≤57.6 Gy
	CI	
	HI	

Abbreviations: CI = conformity index; HI = homogeneity index; PTV = planning target volume.
Apart from obtained values for the constraints, obtained values for parameters in the table without recommended and maximum acceptable values were also used in this study.

For each of the 120 study patients, automated plan generation was performed for the same planning computed tomography and structures as in the clinical plan. Autoplanning was performed with a for breast adapted version of the Guided Planning System[10] in the RayStation TPS, version 11A (RaySearch, Stockholm, Sweden). This autoplanning module was not specifically tuned for generation of highest quality plans for the treatment approaches and traditions in the center where the included patients were treated, as comparison of autoplanning with manual planning was not a study aim (see Introduction section).

Collected data

The labeling of all 240 involved plans as (1) approved without need for further planning to seek improvement or (2) further planning needed was performed by a senior radiation oncologist with more than 20 years of experience in breast cancer treatment (IEO).

The following 24 dosimetric plan parameters were gathered for all 240 plans: $D_{0,03cc}$, $D_{30\%}$, $D_{50\%}$, $D_{95\%}$, conformity index (CI; defined as the ratio between the region of interest volume covered by the 95% isodose and the total patient volume covered by ≥95% of the prescribed dose), and homogeneity index (HI; defined as $D_{95\%}/D_{5\%}$) for the whole breast excluding boost PTV ($PTV_{40.05Gy}$); $D_{0.03cc}$, $D_{5\%}$, $D_{95\%}$, CI, and HI for the boost PTV ($PTV_{48.0Gy}$); V_{20Gy}, V_{8Gy}, and D_{mean} for the heart and D_{mean} and $D_{1\%}$ for the LAD; V_{16Gy}, V_{8Gy}, V_{4Gy}, and D_{mean} for the left lung; V_{4Gy} for the right lung; and $D_{0,03cc}$, $D_{5\%}$, and D_{mean} for CB. See Table 1 for an overview.

Apart from the previously mentioned dosimetric plan parameters, composite dosimetric scores (CPS) were collected for OARs and PTVs, as previously proposed by

IEO investigators.[18] In this scoring system, the involved 5 OARs and 2 PTVs each get a score of 0, 0.5, or 1, depending on the fulfilment of planning constraints reported in Table 1: 1 point was given if all dose constraints were within recommended values, 0.5 point if at least 1 dose constraint was respected, and no points otherwise. Parameters in Table 1 without acceptable values were not considered in this scoring system. Before classifier trainings, the 240 values for each dosimetric feature were first centered around zero by subtracting the mean value, and the values were scaled to unit variance.

The full data set consisted of 240 rows (one for each plan) and 32 columns (24 dosimetric parameters, 7 composite scores, and the clinician's binary score [approved or not]). The Python scikit-learn library[19] was used for all data analyses and model developments.

ML models and training

The investigated 5 dosimetric FS used to train both the RF and LR classifiers (2 × 5 classifiers in total) consisted of the following:

- FS1: 24 dosimetric parameters defined in the Collected Data section
- FS2: 7 CPS defined in Collected Data section
- FS3: 24 differences between dosimetric parameters and their objectives, as indicated in the "Ideal" column of Table 1. If this was missing (eg, left lung D_{mean}), the original value was maintained.
- FS4: FS2 + FS3
- FS5: FS2 + FS1

For each of the 2 × 5 investigated classifiers (RF and LR, both combined with FSi with i = 1-5), model building was performed with nested cross-validation with an outer and an inner loop. The applied procedure is extensively described in Talbot[14] and schematically presented in Fig. 1. Here a brief summary is presented: for the outer loop, the 240 available plans were equally and randomly distributed over 10 folds of 24 plans. Each of the 10 folds then served as a test set for model training based on the remaining (240-24) plans. However, before such a training, an inner-loop 5-fold cross-validation was performed to establish model hyperparameters such as the number and type of trees for RF and solver, penalty, and regularization strength for LR. Inner-loop cross-validations were performed using only the training set of the corresponding outer loop (Fig. 1). For each the 2 × 5 classifiers, the 10 outer-loop models were used to assess the prediction performance. The inner-loop models served only for establishment of model hyperparameters.

The function "GridSearchCV" of the Python scikit-learn library[19] was used in the inner loops to select optimal hyperparameters. For each of the 2 × 5 classifiers, prediction performance was assessed by calculating mean values and standard errors of the accuracy, area under the receiver operating characteristic curve (AUC), and Cohen's kappa coefficient (κ)[20] for the 10 outer-loop models.

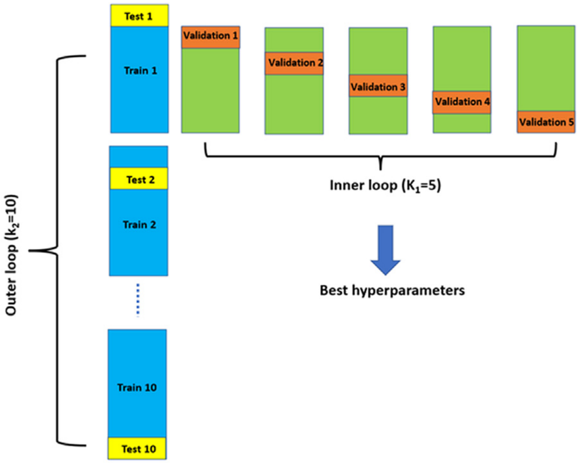

Figure 1 Schematic explanation of the applied nested cross-validation, consisting of 10-fold outer-loop cross-validation and 5-fold inner-loop cross-validation. Each of the 10 outer-loop model buildings is preceded by a paired 5-fold inner-loop cross-validation to establish hyperparameters using only the training patients of the corresponding outer-loop model. Nested cross-validation was performed for each of the 2 × 5 classifiers investigated in this study. For each classifier, the 10 outer-loop models were used to evaluate prediction performance.

Landis and Koch[21] proposed the following classification κ: $\kappa < 0$, agreement "poor"; $0 \leq \kappa \leq 0.2$, agreement "slight"; $0.2 < \kappa \leq 0.4$, agreement "fair"; $0.4 < \kappa \leq 0.6$, agreement "moderate"; $0.6 < \kappa \leq 0.8$, agreement "substantial"; and $0.8 < \kappa \leq 1$, agreement "almost perfect."

For LR, we calculated the Euler number to the power of its coefficient to quantify the importance.[22] For RF classifiers, feature importance was computed as Gini importance or mean decrease in impurity.[23] For each of the 2 × 5 classifiers, the final values of variable importance of included features were calculated as averages of importance values in the 10 outer-loop models. The sum of the importances of all considered features is always 100%.

One-way analysis of variance (ANOVA) tests were used for detecting differences among FS in terms of accuracy, AUC, and κ values, while *t* tests were used for analyzing performance differences between RF and LR classifiers.

Results

The clinician considered all evaluated 240 plans clinically acceptable. Nevertheless, only 92 of 120 clinical plans

Table 2 Accuracy, AUC, and κ parameters for the RF and LR models for the 5 investigated FS

	RF			LR			*P* (Student *t*)		
	Accuracy (%)	AUC	κ	Accuracy (%)	AUC	κ	Accuracy (%)	AUC	κ
FS1	82.5 ± 1.5	0.76 ± 0.03	0.54 ± 0.5	76.7 ± 1.4^{4}	0.75 ± 0.02^{2,4}	0.47 ± 0.03^{4,5}	**.02***	.34	.09
FS2	82.9 ± 1.4	0.76 ± 0.02	0.55 ± 0.04	81.7 ± 1.3	0.85 ± 0.02^{1,3}	0.61 ± 0.03	.22	**.01**†	.09
FS3	85.0 ± 1.1	0.77 ± 0.03	0.59 ± 0.04	75.8 ± 1.7^{4}	0.74 ± 0.02^{2,4,5}	0.45 ± 0.04^{4,5}	**.01***	.18	**.01***
FS4	87.2 ± 2.0	0.80 ± 0.03	0.63 ± 0.05	86.7 ± 2.2^{1,3}	0.86 ± 0.02^{1,3}	0.69 ± 0.04^{1,3}	.43	**.04**†	.19
FS5	83.3 ± 1.8	0.77 ± 0.03	0.56 ± 0.06	83.3 ± 2.5	0.84 ± 0.02^{3}	0.63 ± 0.04^{1,3}	.5	**.04**†	.16
P (ANOVA)	.207	.889	.663	**.005**	**<.001**	**<.001**			

Abbreviations: ANOVA = analysis of variance; AUC = area under the receiver operating characteristic curve; FS = feature set; LR = logistic regression; RF = random forest.

Average values and standard errors were calculated from the 10 folds of the outer loop in the nested cross-validation. The last 3 columns show *P* values for comparisons between RF and LR. The last row shows *P* values for ANOVA tests for models considering FS1 through FS5. Superscript numbers refer to FS that give statistically different results. For example, for FS1, the LR model has an AUC of 0.75 ± 0.02^{2,4}; in this case, superscripts 2 and 4 indicate statistically significant differences for the LR model for FS1 compared with the LR models based on FS2 and FS4, respectively.

In bold the statistically significant values.

* RF is superior.

† LR is superior.

(77%) were approved without further planning, and the remaining 28 not. Of the autoplans, 79 (66%) were judged approved and 41 not.

Model performances in terms of accuracy, AUC, and Cohen's κ are presented and compared in Table 2. Accuracy, AUC, and κ for generated RF/LR models for the most extensive feature set (FS4) were 87.2 ± 2.0/86.7 ± 2.2, 0.80 ± 0.03/0.86 ± 0.02, and 0.63 ± 0.05/0.69 ± 0.04, respectively. Accuracies of 87.2%/86.7% and AUCs of 0.80/0.86 are at the high end compared with many published predictive modeling studies in RT. According to the interpretation by Landis and Koch,[21] κ values of 0.63/0.69 point at "substantial agreement" between clinician plan labeling and ML prediction (see the Methods and Materials section).

The last 3 columns in Table 2 show that performance differences between RF and LR were mostly not statistically significant, depending on considered performance parameter and applied FS. RF had superior accuracy for 2 FS, and one of these also had a superior κ. LR was superior in AUC for 3 FS.

Achieved accuracy, AUC, and κ for created RF classifiers were independent of the FS (columns 2-4 in Table 2, including *P* [ANOVA] in the last row), implying that there was no evidence that adding the FS3 features to FS2 (= FS4; see ML Models and Training section) or adding FS1 to FS2 (= FS5) resulted in better predictions. In contrast, for LR, dependences on applied FS were observed (columns 5-7 in Table 2, including *P* [ANOVA] in last row), with FS4 and FS5 overall performing best.

Figure 2 shows for the evaluated PTVs and OARs, summed importances of the corresponding features for the 5 investigated FS (left panel: RF, right panel: LR). Both for RF and LR, $PTV_{40.05Gy}$ was undoubtedly the most important structure for the predictions, independent of the applied FS. The right lung was always of minor importance, and the most important OARs were heart and LAD for RF and LR classifiers, respectively. Figure 3

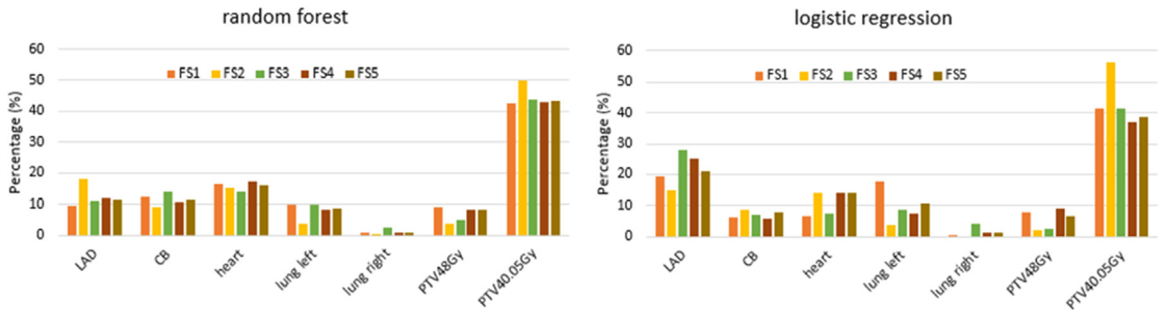

Figure 2 Importances for all feature sets (FS1-FS5) for all considered structures (organs at risk and planning target volumes). For each feature set, the values for the 7 structures add up to 100%. For each structure, the bar for each feature set represents the sum of the importances of all features related to that structure. *Abbreviations*: CB = contralateral breast; LAD = left anterior descending coronary artery; PTV = planning target volume.

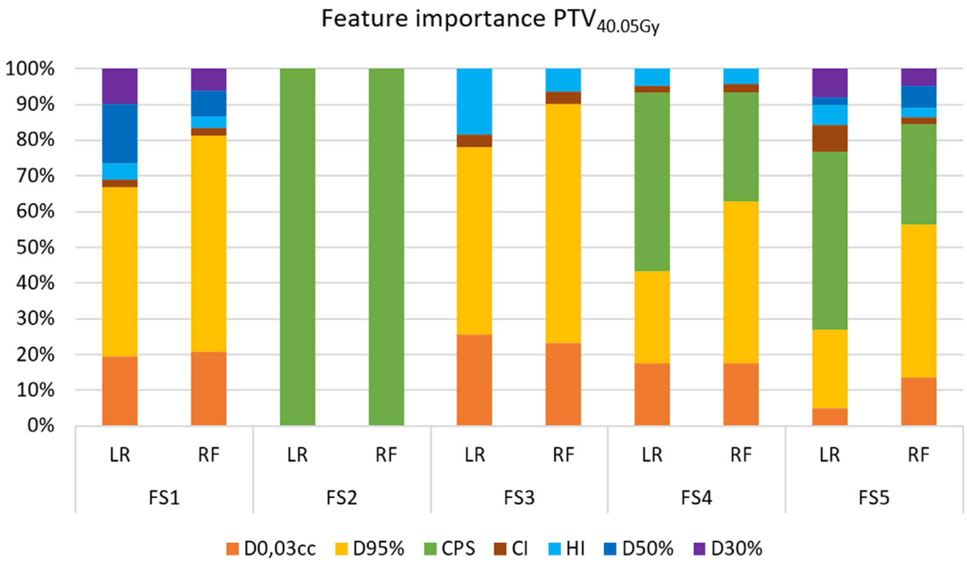

Figure 3 Relative importance of various $PTV_{40.05Gy}$ features. *Abbreviations*: CI = conformity index; CPS = composite score; HI = homogeneity index; LR = logistic regression; PTV = planning target volume; RF = random forest.

shows importances of various $PTV_{40.05Gy}$ features: for FS1 and FS3, which do not contain the CPS, it is clearly seen that $D_{95\%}$ is the parameter that has the highest importance in predictions, and for FS4 and FS5 with LR model the CPS and with RF model the $D_{95\%}$ are respectively the features of greater importance.

Discussion

In the complex landscape of large amounts of data, ML (including deep learning) offers unique opportunities for improving the overall quality and efficiency of the modern RT workflow.[24-27] The aim of our study was to investigate whether ML models could become useful for aiding treatment planners to present only treatment plans to clinicians that have a high probability to be approved without further planning. The applied data set consisted of 240 treatment plans for left-sided WBRT with boost, each of them retrospectively labeled by a clinician as either “approved without further planning to improve” or “further planning needed.” In total, 2 × 5 classifiers were investigated; RF and constraint LR, both trained with 5 different sets of dosimetric plan features. For a given treatment plan, each of the 2 × 5 classifiers predicted the probability that the clinician would approve the plan without further planning. For plans labeled “approved without further planning to improve,” a probability >.5 was considered as a correct prediction. Likewise, for plans with a label “further planning needed,” a probability <.5 was considered correct. The use of 5 different FS allowed us to investigate the sensitivity of RF and LR for the choice of applied dosimetric features. FS1 and FS3 both consisted of 24 dosimetric parameters that could be directly calculated from the dose distributions. The much smaller FS2 (7 parameters) contained for each of the 7 involved anatomic structures a composite score that was derived from related dosimetric parameters, as previously proposed.[18] FS4 and FS5 were the largest FS, consisting of FS2 + FS3 and FS2 + FS1, respectively. For FS4, accuracy, AUC, and Cohen’s κ for generated RF/LR models for prediction of approval without further planning were rather high: 87.2 ± 2.0/86.7 ± 2.2, 0.80 ± 0.03/0.86 ± 0.02, and 0.63 ± 0.05/0.69 ± 0.04, respectively. RF performance was basically independent of the applied FS (Table 2), meaning that FS2 with only 7 features performed as well as FS1 and FS3 with 24 features and FS4 and FS5 with 31 features. For LR, a dependency on FS was observed, with the large FS4 and FS5 overall performing the best. The possibility of using nonlinear combinations of available dosimetric features in RF modeling could make up for reduced availability of dosimetric features in FS2.

Clinicians’ plan evaluations are not only based on plan parameters but consider also the full 3D dose distribution. This study shows that not explicitly considering the full 3D dose in the 2 × 5 investigated classifiers could still result in high-quality predictions of clinicians’ plan evaluations.

As mentioned previously, all 240 study plans were retrospectively labeled by the clinician as “approved without further planning” or “further planning needed.” Apart from this labeling, the clinician also assessed plan acceptability. Although only 71.5% of plans were labeled as “approved without further planning,” the clinician found 100% of plans in principle acceptable for treatment. Apparently, for a large number of plans the clinician had a wish to further explore plan improvement even though the plan was in principle acceptable. This reflects the complex decision making that was modeled in this article; the label “further planning needed" was not related to unacceptable constraint violations but to more subtle desires for plan improvement.

For all investigated 2 × 5 classifiers, $PTV_{40.05Gy}$ was by far the most important anatomic structure for predictions, reflecting the importance given to it by the clinician (Fig. 2), with $D_{95\%}$ as the most important parameter for most classifiers having $PTV_{40.05Gy}$ $D_{95\%}$ as feature (Fig. 3).

In this study, all 240 available labeled treatment plans could be used for training, validation, and testing (classifier performance assessment) due to the applied nested cross-validation (Talbot,[14] Fig. 1). With this procedure, inner-loop cross-validation was used for establishment of hyper parameters, to be used for model trainings in the outer-loop cross-validation.

A limitation of this study is that the analyses were performed for a single clinician. Generalizability of these prediction models for use by more clinicians is a topic of future research. The endeavor of developing a single model for all clinicians in the center could result in higher consistency of the treatments delivered in the study center. Another limitation is the lack of nondosimetric patient data in the performed analyses, including age, performance status, previous or concomitant treatments, surgery results, and comorbidities. Future investigations will include such factors that could further enhance the reliability of the predictive models.

Conclusion

We have investigated several ML approaches for prediction of clinician approval of treatment plans for left-sided WBRT plus boost based on plan dosimetry. Results are encouraging for future workflows in which treatment planners will only present treatment plans to treating clinicians if they have a high probability of being directly approved, that is, without an additional round of planning and plan evaluation.

Acknowledgments

The authors thank the editor and reviewers for providing very helpful inputs that served to better clarify some key points in the article.

References

1. Pignol JP, Olivotto I, Rakovitch E, et al. A multicenter randomized trial of breast intensity-modulated radiation therapy to reduce acute radiation dermatitis. *J Clin Oncol*. 2008;26:2085-2092.
2. Mukesh MB, Barnett GC, Wilkinson JS, et al. Randomized controlled trial of intensity-modulated radiotherapy for early breast cancer: 5-year results confirm superior overall cosmesis. *J Clin Oncol*. 2013;31:4488-4895.
3. Franco P, Catuzzo P, Cante D, et al. TomoDirect: An efficient means to deliver radiation at static angles with tomotherapy. *Tumori*. 2011;97:498-502.
4. Murai T, Shibamoto Y, Manabe Y, et al. Intensity-modulated radiation therapy using static ports of tomotherapy (TomoDirect): Comparison with the TomoHelical mode. *Radiat Oncol*. 2013;8:68.
5. Franco P, Zeverino M, Migliaccio F, et al. Intensity-modulated adjuvant whole breast radiation delivered with static angle tomotherapy (TomoDirect): A prospective case series. *J Cancer Res Clin Oncol*. 2013;139:1927-1936.
6. Dicuonzo S, Leonardi MC, Raimondi S, et al. Acute and intermediate toxicity of 3-week radiotherapy with simultaneous integrated boost using TomoDirect: Prospective series of 287 early breast cancer patients. *Clin Transl Oncol*. 2021;23:1415-1428.
7. Kisling K, Johnson JL, Simonds H, et al. A risk assessment of automated treatment planning and recommendations for clinical deployment. *Med Phys*. 2019;46:2567-2574.
8. Marrazzo L, Meattini I, Arilli C, et al. Auto-planning for VMAT accelerated partial breast irradiation. *Radiother Oncol*. 2019;132:85-92.
9. Redapi L, Rossi L, Marrazzo L, et al. Comparison of volumetric modulated arc therapy and intensity-modulated radiotherapy for left-sided whole-breast irradiation using automated planning. *Strahlenther Onkol*. 2022;198:236-246.
10. Fiandra C, Rossi L, Alparone A, et al. Automatic genetic planning for volumetric modulated arc therapy: A large multi-centre validation for prostate cancer. *Radiother Oncol*. 2020;148:126-132.
11. Rossi L, Sharfo AW, Aluwini S, et al. First fully automated planning solution for robotic radiosurgery: Comparison with automatically planned volumetric arc therapy for prostate cancer. *Acta Oncol*. 2018;57:1490-1498.
12. Heijmen B, Voet P, Fransen D, et al. Fully automated, multi-criterial planning for volumetric modulated arc therapy: An international multi-center validation for prostate cancer. *Radiother Oncol*. 2018;128:343-348.
13. Cagni E, Botti A, Rossi L, et al. Variations in head and neck treatment plan quality assessment among radiation oncologists and medical physicists in a single radiotherapy department. *Front Oncol*. 2021;11: 706034.
14. Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.
15. Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early-stage breast cancer. *Radiother Oncol*. 2015;114:3-10.
16. Duane F, Aznar MC, Bartlett F, et al. A cardiac contouring atlas for radiotherapy. *Radiother Oncol*. 2017;122:416-422.
17. Radiation Therapy Oncology Group (RTOG) 1005. A phase III trial of accelerated whole breast irradiation with hypofractionation plus concurrent boost versus standard whole breast irradiation plus sequential boost for early-stage breast cancer. Available at: https://www.nrgoncology.org/Clinical-Trials/Protocol/rtog-1005.
18. Orecchia R, Rojas DP, Cattani F, et al. Hypofractionated postmastectomy radiotherapy with helical tomotherapy in patients with immediate breast reconstruction: dosimetric results and acute/intermediate toxicity evaluation. *Med Oncol*. 2018;35(3):39.
19. Pedregosa F, Varoquaux G, Gramfort, et al., et al. Scikit-learn: Machine learning in Python. *JMLR*. 2011;12:2825-2830.
20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurement*. 1960;20:37-46.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
22. König G, Molnar C, Bischl B, Grosse-Wentrup M. *Relative feature importance*. Boston: IEEE Xplore; 2020:623-630.
23. Qi Y. Random forest for bioinformatics. In: Zhang C, ed. *Ensemble Machine Learning*. New York, NY: Springer; 2012:307-323. Ma, YQ.
24. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273-297.

25. Sadeghnejad Barkousaraie AS, Ogunmolu O, Jiang S, Nguyen D. A fast deep learning approach for beam orientation optimization for prostate cancer treated with intensity-modulated radiation therapy. *Med Phys*. 2019;47:880-897.

26. Granville DA, Sutherland JG, Belec JG, La Russa DJ. Predicting VMAT patient-specific QA results using a support vector classifier trained on treatment plan characteristics and linac QC metrics. *Phys Med Biol*. 2019;64: 095017.

27. Li J, Wang L, Zhang X, Liu L, et al. Machine learning for patient-specific quality assurance of VMAT: Prediction and classification accuracy. *Int J Radiat Oncol Biol Phys*. 2019;105:893-902.

28. Hussein M, Heijmen BJM, Verellen D, Nisbet A. Automation in intensity modulated radiotherapy treatment planning: A review of recent innovations. *Br J Radiol*. 2018;91: 20180270.