RESEARCH ARTICLE

Artificial Intelligence-Assisted Peer Review in Radiation Oncology

Renee F. Cattell, PhD;[1]* Jinkoo Kim, PhD; Ewa Zabrocka, MD;[1,2] Xin Qian, PhD; Brian O'Grady, BA;[1] Stephanie Butler, BS; Todd Yoder, MS, CMD;[1,3] Kartik Mani, MD, PhD; Mark Ashamalla, MD; Samuel Ryu, MD

Abstract

Objective Peer review is an essential part of the patient treatment process that examines and, where necessary, recommends revisions to clinical data, therapeutic parameters, and potential alternative approaches to treatment. Our hypothesis is that artificial intelligence (AI) and machine language technologies can enhance peer-review efficacy by screening cases for potential treatment interruptions caused by re-planning and treatment cessation.

Materials and Methods Fifty-five features of clinical and therapeutic parameters from 3881 radiotherapy patients (7142 plans) treated from 2014 to 2021 were used as input for two AI models: a multivariable least absolute shrinkage and selection operator (LASSO) logistic regression model and a pattern recognition feed-forward neural network (NN). The dataset was split into 70% training and 30% testing, with the training set divided into five groups for cross-validation. Analysis was performed on the full cohort and on subsets based on treatment site. Performance metrics of accuracy, sensitivity, and specificity were calculated.

Results Overall, 8.1% of all cases had treatment interruptions, most commonly in the head and neck region compared to other sites (19% vs 6%-9%, $P<.01$). For the LASSO model, test set sensitivity, specificity, and accuracy ranged from 37%-70%, 59%-78%, and 60%-76%, respectively, with higher specificity than sensitivity for site subsets. For the NN model, test set sensitivity, specificity, and accuracy ranged from 41%-68%, 53%-79%, and 53%-78%, respectively. Both models demonstrated the highest accuracy in the brain subset. For the full cohort, NN accuracy (58%) was similar to LASSO (60%). The largest accuracy differences between LASSO and NN were in the lung/breast/chest (LASSO: 71% vs NN: 57%) and spine/extremity (LASSO: 66% vs NN: 54%) subsets.

Conclusion Our results provide proof-of-concept that AI- and ML-based technologies have potential as screening tools to aid peer review in radiation oncology. Early identification of patients at risk for radiation therapy interruptions using these tools could translate into higher treatment completion rates. The study is being continued to include more clinical features and to optimize model hyperparameters.

Keywords: peer review, artificial intelligence, machine learning, radiation oncology

Affiliations: [1]Department of Radiation Oncology, Stony Brook University Hospital, Stony Brook, NY. [2]Department of Radiation Oncology, Anchorage and Valley Radiation Therapy Centers, Anchorage, AK. [3]Department of Radiation Oncology, NYU Langone Health, New York, NY.
Corresponding author: *Renee F. Cattell, PhD, Stony Brook University Hospital, 101 Nicolls Rd, Stony Brook, NY 11794. (renee.cattell@stonybrookmedicine.edu)
Disclosures: The authors have no conflicts of interest to disclose. None of the authors received outside funding for the production of this original manuscript and no part of this article has been previously published elsewhere.
Prior Publication/Presentation: Some of the material in this article was published as an abstract in the following supplement for the ASTRO 2022 Annual Meeting: Cattell R, Ashamalla M, Kim J, et al. Artificial intelligence–assisted peer review in radiation oncology. *Int J Radiat Oncol Biol Phys*. 2022;114(3)(suppl):e471. doi:10.1016/j.ijrobp.2022.07.1726.

Published: July 1, 2025. https://doi.org/10.37549/ARO-D-24-00025

Introduction

Peer review in radiation oncology is essential for the safe and efficient delivery of radiation treatments;[1] however, little guidance and limited research exist regarding the frequency, mechanisms, and metrics of peer review from professional organizations.[1-4] Peer review addresses patient characteristics and clinical oncological information, planned radiotherapeutic parameters, and potential treatment-related variations. In a report series commissioned by the American Society for Radiation Oncology, seven main items are currently examined by peer reviewers.[1] These are (1) the decision to include radiation as part of treatment, (2) the general radiation treatment approach, (3) the target definition, (4) normal tissue image segmentation, (5) the planning directive, (6) technical plan quality, and (7) treatment delivery.

Many of these factors are indicative of cases requiring re-plans or cases that may have treatment interruptions. Although peer reviews should be conducted prior to the start of treatment to ensure safety and quality, they are typically performed either immediately before or after treatment has begun. Thus, there is minimal time to adjust therapy based on the recommendations of peer reviewers. Plan changes based on peer review discussions are not uncommon. Hoopes et al reported that 90% of physicians have changed their radiation plans because of peer review.[3] Other studies have shown that up to 10% of cases have been recommended for plan modifications based on peer review.[1,3] Studies have also demonstrated that prolonged radiation treatment times, for various reasons, correlated with reduced local disease control and worse overall survival.[5-7] Therefore, timely, efficient peer review may greatly improve treatment quality and, ultimately, enhance patient safety.

However, peer review has some practical difficulties, especially when the process must be coordinated among multiple facilities. While a fully integrated approach holds promise for improving the quality, safety, and value of cancer care, in reality, the process is often disjointed owing to differing provider schedules, caseloads, and expertise. These opportunities to optimize peer review are further highlighted by the fact that different approaches are taken by academic centers and community cancer centers operating within the same network.[8] Practically speaking, however, caseload, time, department resources, staffing, and increasing complexity in treatment techniques limit thorough peer review.[9] Thus, detecting cases that may experience treatment interruptions due to undesirable effects is not easy.

A pan-Canadian survey by Caissie et al demonstrated that barriers to peer review, including time constraints (27%) and radiation oncologist availability (34%), caused half of all programs surveyed to conduct peer review after the start of treatment.[10] Therefore, automating peer review, even if only partially, may significantly streamline the process.

Artificial intelligence (AI) tools can potentially be used to assist in peer review. Studies have used machine learning (ML) techniques to identify high-risk patients for detailed clinical evaluation during radiation and chemoradiation.[11] To identify potential treatment interruptions or unusual side effects resulting from suboptimal treatment plans, we hypothesize that complex, nonlinear relationships exist between different variables, including clinical characteristics and plan parameters. Our approach was to use AI and ML to uncover complex interactions among variables and to supplement clinical knowledge with treatment and plan parameters. By using AI, our intent was to screen treatment plans for clinical and radiotherapeutic factors that could lead to treatment interruptions or toxicity and submit the plans for more detailed inspection to help expedite peer review.

Methods and Materials

Data Collection

This retrospective study was approved by the Institutional Review Board. Clinical data and radiation therapy plan parameters from 3881 patient records (7142 plans) were retrospectively extracted from electronic medical records and treatment planning systems from January 2014 to March 2021 for patients older than 18 years. The study inclusion criteria consisted of patients who received external beam photon radiation therapy with either curative or palliative intent for initial treatment and/or re-treatment. This study excluded patients who underwent brachytherapy and electron therapy. Included patients were further divided into subsets based on bodily treatment sites (**Table 1**). Comparisons were made using two-tailed *t* tests with unequal variance.

The aim of this study was to identify patients likely to experience treatment interruptions (eg, changes in target volume, changes in prescription dose) or complications (eg, toxicity). As a surrogate for this outcome, plans with treatment interruptions (remaining fractions) were designated as abnormal cases,

Table 1. Distribution of Plans With and Without Treatment Interruptions Across Separate Subset Groups. "All" Indicates the Full Cohort Before Separated into Subsets Based on Treatment Site

TREATMENT SITE	TOTAL NUMBER OF PLANS	PLANS WITH TREATMENT INTERRUPTIONS	PLANS WITHOUT TREATMENT INTERRUPTIONS
All	7142	579 (8.1%)	6563 (91.9%)
Lung, breast, and chest	3329	228 (6.8%)	3101 (93.2%)
Pelvis and prostate	1077	90 (8.4%)	987 (91.6%)
Spine and extremity	983	92 (9.4%)	891 (90.6%)
Brain	1285	81 (6.3%)	1204 (93.7%)
Head and neck	468	88 (18.8%)	380 (81.2%)

whereas radiation therapy plans with no remaining fractions (ie, no interruptions) were considered to be normal cases. We defined remaining fractions as any plan that was not completed; we did not separate re-planned cases versus discontinued cases.

Most interruptions are the result of toxicity or treatment response based on radiosensitivity and tumor histology. This surrogate, although imperfect, was easily translated into an instance that could be extracted automatically from the patient records. We note that the specific reason(s) for treatment interruption were not included in this study, as our intention was to identify potential cases and prevent treatment modification. Various clinical factors and therapeutic technical parameters were collected for logistic regression analysis and developing neural network (NN) models. The input factors are listed in **Table 2**.

Predictive Modeling

The two classification models used in this study were a multivariable logistic regression with least absolute shrinkage and selection operator (LASSO) and a pattern recognition feed forward NN. For LASSO, the maximum number of non-zero coefficients was set to 10. The NN model had one hidden layer with 10 neurons, and the weights were initialized with random seeds. The scaled conjugate gradient algorithm was used for training, and the mean absolute error was the cost function. Regularization was performed using error weights to prevent overfitting of the NN model due to the unbalanced dataset; entry samples with treatment interruptions had two times the weight of those without treatment interruption for training.

For both techniques, the dataset was first split into a 70% training set and a 30% testing set. The training set was normalized with z-score normalization, with the same center and standard deviation normalization applied to the testing set. The training set was further divided into five folds for cross-validation. Within each cross-validation, the minority class was oversampled using adaptive synthetic sampling.[12] The validation fold was not oversampled. The testing set was not oversampled or included in the training of the algorithms. The number of input features to each model was 55. MATLAB (2022a, The MathWorks, Inc., Natick Massachusetts, United States) was used for development, training, and evaluation of both models.

Predictive Performance Evaluation

The full cohort and subsets were analyzed based on treatment site. The optimal operating point threshold of the receiver-operating characteristic curve was determined only from the training set. This same threshold was applied to the validation and testing sets. Sensitivity, specificity, and accuracy were calculated. The model with the greatest validation accuracy across the five-fold cross-validation was used to predict the independent testing set to assess predictive performance. Comparisons were made using two-tailed paired *t* tests.

RESULTS

Description of Cohort

A total of 3881 patients with 7142 plans made up the full cohort. Of the cohort, 58.1% were between 50 and 74 years, 33.1% were >75 years, and 8.8% were between 19 and 49 years. There were more females than males (59.6% vs 40.4%). The most common treatment sites were the breasts (30.2%), brain (18.0%), and lungs (13.0%). Most patients were treated with primary definitive intent (74.6%) versus treatment of metastatic disease (25.4%). **Table 1** summarizes the study cohort. As shown, 8.1% of all plans experienced treatment interruptions. The head-and-neck subset had the largest percentage of treatment interruptions (18.8%), while the other subsets ranged from 6.3% to 9.4% ($P < .01$).

Testing Set Breakdown

The model with the highest accuracy from the five-fold cross-validation was selected as the

Table 2. Input Features for the LASSO and NN Models

CLINICAL FEATURES	PLAN PARAMETER FEATURES
Patient sex (male, female)	Number of beams
Patient age group in years (≥75, 50-74,19-49)	Type of plan (IMRT, RapidArc, other)
Site	Monitor units (total, average, minimum, maximum)
Primary or metastasis	Source to skin distance (SSD; average, minimum, maximum)
Inpatient status	Table tolerance
Personal history of cancer	Gantry angle (minimum, maximum)
Family history of cancer	Bolus
Tobacco use	Gating
Smoker	Collimator angle (average, minimum, maximum)
Second hand smoke	Couch lateral (maximum)
Obesity	Couch longitudinal (maximum)
Alcohol	Couch vertical (maximum)
Human papillomavirus (HPV)	Couch angle (average, minimum, maximum)
Immunosuppression	Field size (average, minimum, maximum)
Immunodeficiency	Isocenter (*X*, *Y*, *Z*)
Neutropenia	Maximum beam energy
Anemia	Dose per fraction
Human immunodeficiency virus (HIV)	
Hepatitis	
Heart disease	
Diabetes	
Hypertension	
Hyperlipidemia	
Failure to thrive	

Abbreviations: LASSO, least absolute shrinkage and selection operator; NN, neural network; IMRT, intensity-modulated radiation therapy.

predictive model on an independent testing set. The breakdown of plans with and without treatment interruptions in the selected folds are shown in **Table 3**. If the same fold was selected for the LASSO and NN models, the split was the same for training and validation sets. If a different fold was selected, the number of positive samples may be slightly different, owing to randomization during the splitting process. For the testing set, the split was the same for LASSO and NN because this was a hold-out set and not involved in the training process.

LASSO Performance

The performance metrics for the LASSO model are shown in **Figure 1** and Supplemental Table A (**www.appliedradiationoncology.com**). For the training set, the average sensitivity, specificity, and accuracy ranged from 69% to 92%, 52% to 67%, and 67% to 79%, respectively. Sensitivity was significantly higher than specificity for the full cohort, the spine/extremity subset, and the brain subset (P<.05). For the validation set, the average sensitivity, specificity, and accuracy ranged from 42% to 82%, 49% to 64%, and 51% to 64%, respectively. Sensitivity remained significantly higher than specificity only for the full cohort (P<.05).

For the independent testing set, the sensitivity, specificity, and accuracy ranged from 37% to 70%, 59% to 78%, and 60% to 76%, respectively. Except for the full cohort, specificity was higher than sensitivity. The brain subset had the highest accuracy (76.1%). Overall, the higher specificity than sensitivity on the independent testing set indicated that the model was better able to predict true negatives (those without treatment interruptions) than true positives (those with treatment interruptions).

For the LASSO model, **Table 4** shows the features selected by the algorithm for the prediction, which included the components of clinical features and plan parameters.

Neural Network Performance

The performance metrics for the NN model are shown in **Figure 2** and Supplemental Table B (**www.appliedradiationoncology.com**). For the training set, the average sensitivity, specificity, and accuracy ranged from 79% to 99%, 34% to 68%, and 59% to 83%, respectively. Sensitivity was significantly higher than specificity (P<.05) for all subsets except the pelvis/prostate subset (P=.15). For the validation set, average sensitivity, specificity, and accuracy ranged from 53% to 77%, 35% to 68%, and 38% to 67%, respectively. We observed a similar trend in the validation set as in the training set; sensitivity in the validation set was generally higher than specificity,

Table 3. Distribution of Plans With and Without Treatment Interruptions Across Separate Subset Groups from the Selected Cross-Validation Fold With the Highest Validation Set Sensitivity. Numbers Shown Indicate the Number of Plans With Treatment Interruptions Divided by the Total Number of Plans. "All" Indicates the Full Cohort Before Separated into Subsets Based on Treatment Site

	TRAINING SET		VALIDATION SET		TESTING SET	
	LASSO	NN	LASSO	NN	LASSO	NN
All	315/3999 (7.9%)	297/3999 (7.4%)	76/1001 (7.6%)	94/1001 (9.4%)	188/2142 (8.8%)	188/2142 (8.8%)
Lung, breast, and chest	129/1864 (6.9%)	141/1864 (7.6%)	42/467 (9.0%)	30/467 (6.4%)	57/998 (5.7%)	57/998 (5.7%)
Pelvis and prostate	50/603 (8.3%)	51/604 (8.4%)	13/152 (8.6%)	12/151 (7.9%)	27/322 (8.4%)	27/322 (8.4%)
Spine and extremity	56/550 (10.2%)	56/550 (10.2%)	11/139 (7.9%)	11/139 (7.9%)	25/294 (8.5%)	25/294 (8.5%)
Brain	32/719 (4.5%)	32/719 (4.5%)	13/181 (7.2%)	13/181 (7.2%)	36/385 (9.4%)	36/385 (9.4%)
Head and neck	47/262 (17.9%)	50/262 (19.1%)	14/67 (20.9%)	11/67 (16.4%)	27/139 (19.4%)	27/139 (19.4%)

Abbreviations: LASSO, least absolute shrinkage and selection operator; NN, neural network

although it was only statistically significant in the spine/extremity subset ($P<.05$). For the testing set, the sensitivity, specificity, and accuracy ranged from 41% to 68%, 53% to 79%, and 53% to 78%, respectively. Overall, accuracy was highest in the brain subset (78%).

DISCUSSION

These study results provide proof of concept that AI can be a reliable screening tool in the peer review process to help identify cases early on that may cause treatment interruption or major changes in treatment course. We found that logistic regression and NN-based models had some predictive power in recognizing cases that would experience treatment interruptions. This could be useful in identifying cases that will require a replan for various reasons, such as a patient who experiences toxicity or has a dramatic change in target volume. Cases identified as "high risk" for interruptions could be highlighted in peer review for a more in-depth evaluation.

Our study demonstrates the potential of AI in radiation oncology peer review to prospectively identify treatment interruption. A simple example of one potential software display in a peer-review setting is shown in **Figure 3**. Each patient could be assigned a "score" associated with treatment interruption risk that could help prioritize cases for discussion. The software might also display the clinical or plan features flagged for that specific case. Links directly to the plan and other relevant clinical documents could facilitate quick reference during discussions.

In radiation oncology, every patient receives a personalized plan chosen by their physician that best fits their characteristics, such as their diagnosis and performance status. However, for reasons that are sometimes unknown, a patient may require a change or break in treatment, which can be detrimental to their oncologic outcome. In patients with head-and-neck cancer, the hazard rate of death increased 4.2% for each additional day needed to finish radiation therapy beyond 8 weeks.[13] Even small disruptions in radiation therapy can have negative consequences in gynecologic patients. Lanciano et al reported a 7.7% reduction in 4-year survival when the radiation therapy course was >10 weeks compared with 8.0-9.9 weeks.[7]

The outcome analyzed in our study was based on whether or not a plan had remaining fractions. Although remaining fractions are not always indicative of treatment toxicity, they could be a major contributor. Several studies have looked at the ability of ML models to predict toxicity using clinical and dose factors and/or radiomic features.[14,15] Reddy et al used random forest, gradient-boosted decision tree, and logistic regression models with input of clinical and treatment variables to predict breast cancer treatment toxicity and achieved an area under the curve (AUC) ranging from 0.56 to 0.85.[16] Das et al created a fusion of different non-linear multivariate models (decision trees, NNs, support vector machines, and self-organizing maps) with input of dose and non-dose patient variables to predict

Figure 1. Predictive performance of the least absolute shrinkage and selection operator (LASSO) model for (A) training, (B) validation, and (C) testing datasets. For the training and validation datasets, it is the average across five-fold cross-validation. "All" indicates the full cohort before separation into subsets based on treatment site and error bars indicate standard deviation.

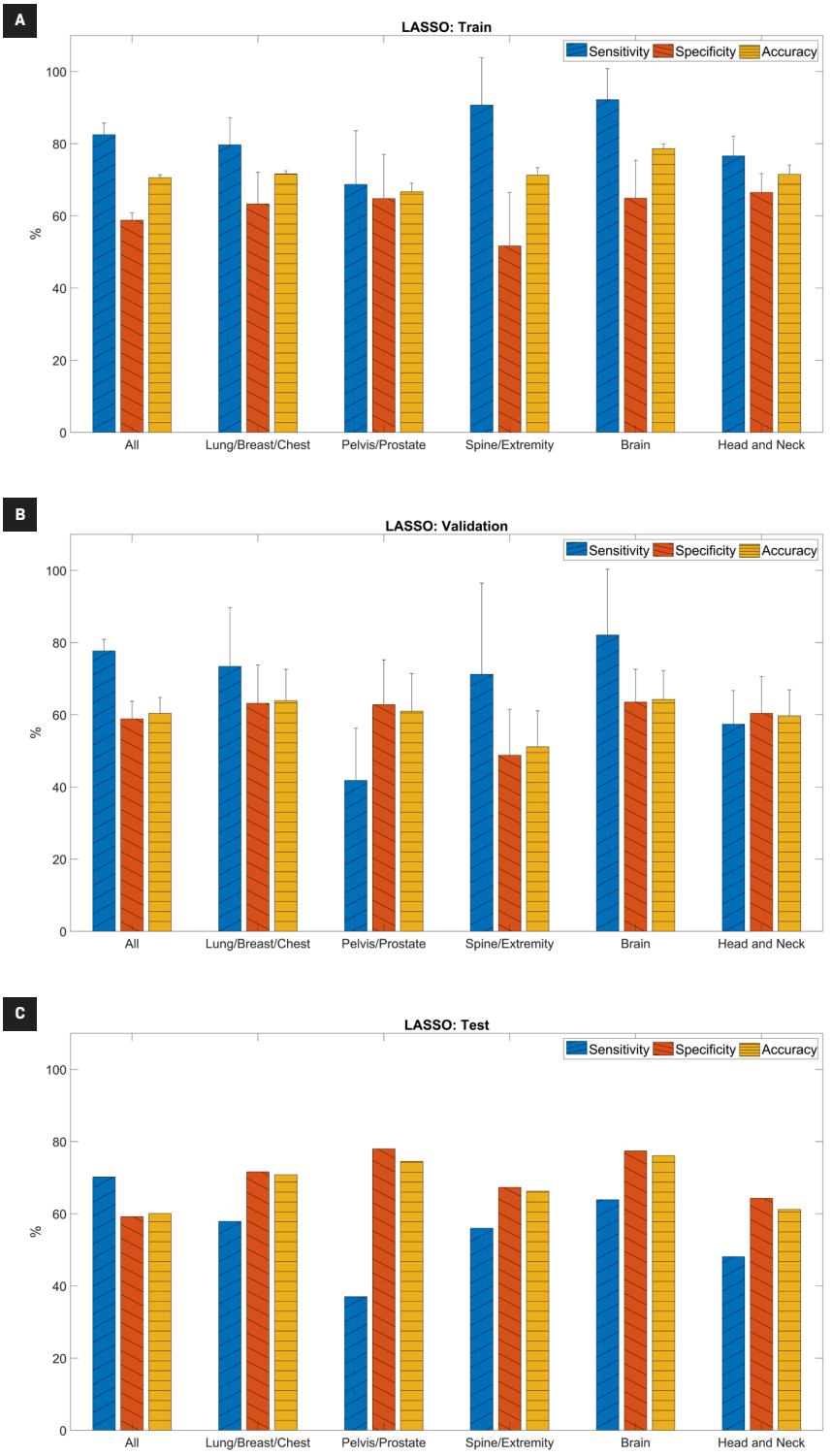


Table 4. Features Selected by Logistic Regression Least Absolute Shrinkage and Selection Operator (LASSO) Model for Each Subset Group. "All" Indicates the Full Cohort Before Separation into Subsets Based on Treatment Site. Isocenter Location *X*, *Y*, and *Z* Refer to Lateral, Anterior/Posterior, and Superior/Inferior Directions, Respectively

SITE	CLINICAL FEATURES	PLAN FEATURES
All	• Sex • Inpatient status • Personal history of cancer	• Monitor units (max) • Source to skin distance (max) • Gating • Collimator rotation (max) • Field size (average) • Field size (max) • Dose per fraction
Lung, breast and chest	• Inpatient status • Personal history of cancer	• Type of plan • Monitor units (average) • Source to skin distance (average) • Source to skin distance (max) • Bolus • Gating • Field size (min) • Field size (max)
Pelvis and prostate	• Anemia • Diabetes	• Type of plan • Couch longitudinal (max) • Field size (max) • Dose per fraction
Spine and extremity	• Patient age group • Inpatient status • Tobacco	• Monitor units (total) • Source to skin distance (min) • Couch longitudinal (max) • Field size (max) • Isocenter (*X*) • Energy (max) • Dose per fraction
Brain	• Inpatient status • Immunosuppression • Anemia	• Monitor units (average) • Tolerance table • Gantry angle (max) • Bolus • Couch vertical (max) • Isocenter (*Z*) • Dose per fraction
Head and neck	• Patient age group • Inpatient status • Tobacco • Obesity • Hypertension	• Monitor units (average) • Monitor units (min) • Tolerance table • Field size (average)

radiation-induced pneumonitis in lung cancer patients with an AUC of 0.79.[17]

The test set accuracy in our study ranged from 57% to 71% for the lung/breast/chest subset. We grouped lung and breast into the same subset due to data size limitations; thus, the predictive accuracy of our models may be improved by further separation.

Similar to their study on breast cancer patients, Reddy et al used random forest, gradient-boosted decision tree, and logistic regression models with clinical and treatment parameters to predict head-and-neck cancer treatment toxicity and achieved an AUC of 0.64-0.76 in their validation set.[18] Jiang et al used three supervised learning methods (ridge logistic regression, lasso logistic regression, and random forest) to predict xerostomia, resulting in an AUC of 0.7.[19]

Our head-and-neck subset came in lower at 53%-61% test set accuracy. We did not directly analyze toxicity because we predicted whether there were remaining fractions in each plan. Remaining fractions could also be caused by re-planning owing to tumor shrinkage; this may explain the reduced predictive value of our study compared to others looking solely at toxicity. Carrara et al applied the artificial NN approach with five input variables relating to patient dose, history, and therapy to predict toxicity after high-dose prostate cancer radiation therapy, achieving an AUC of 0.78.[20] Pella et al used support vector machines and neural network-based algorithms to predict acute toxicity of the bladder and rectum due to prostate irradiation, resulting in overall accuracy similar in both models at an AUC of 0.7.[21] The accuracy of our testing set in the prostate/pelvis subset, at 71%-75%, is similar.

Figure 2. Predictive performance of the neural network (NN) model for (A) training set, (B) validation set, and (C) testing set. For the training and validation sets, it is the average across five-fold cross-validation. "All" indicates the full cohort before separation into subsets based on treatment site and error bars indicate standard deviation.

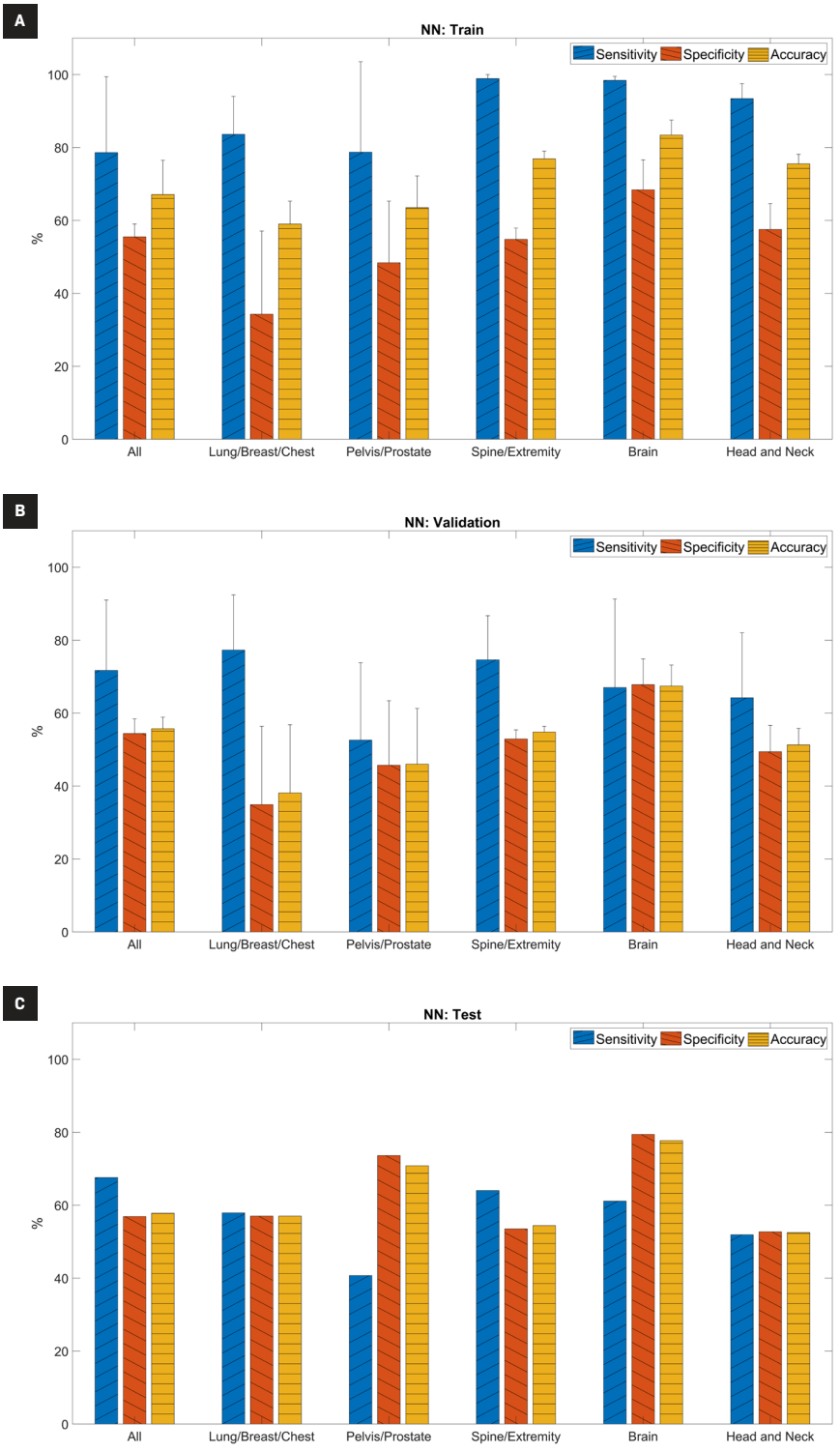
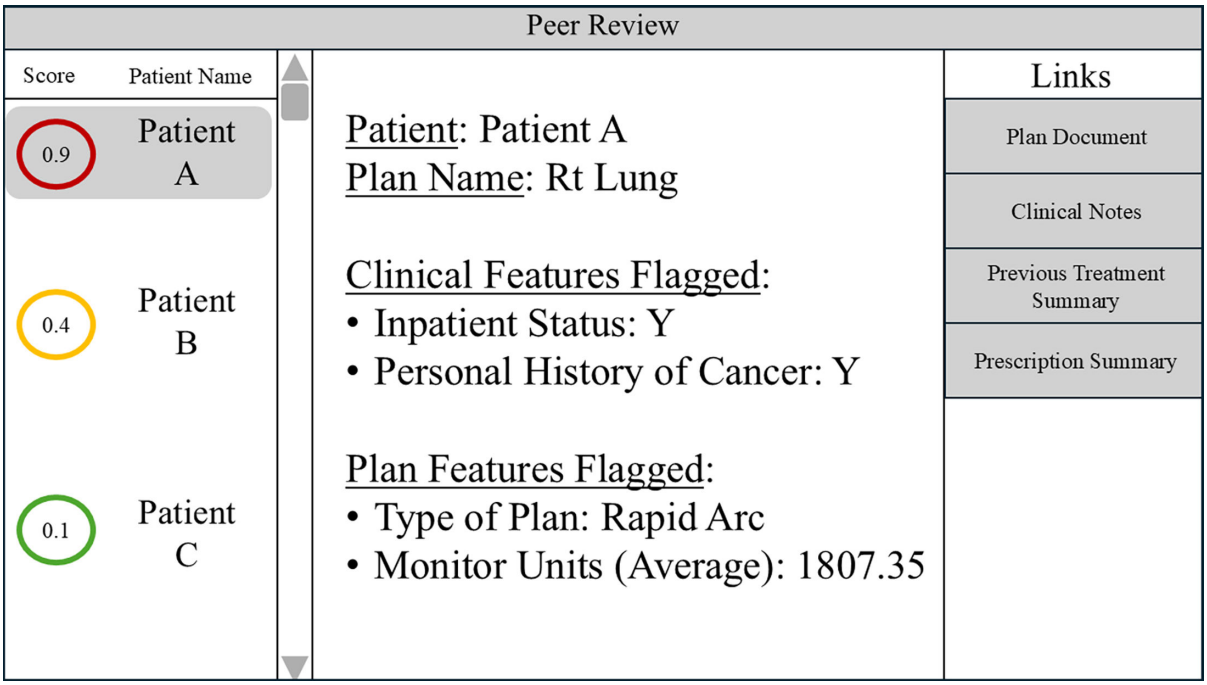


Figure 3. An example of an application interface that can be displayed during peer review. The application incorporates the risk of treatment interruption score, highlighting features flagged for review and quick access links to patient data.



Although toxicity is one cause of treatment interruption, physician input/suggestions can also cause interruptions if the case is not presented before treatment and/or some information is not assessed during peer review. The traditional approach to peer review is based on "chart rounds," where the physicians, physicists, dosimetrists, and therapists review details of each case (eg, clinical history, treatment technique, prescription dose, treatment plan, and patient setup). The average amount of time spent on each patient during peer review was reported to range between 1 and 4 minutes.[2,22]

Therefore, reviewing the more technical aspects of treatment delivery (eg, monitor units and couch/collimator/gantry parameters), which may provide additional information, may not be feasible owing to time constraints. As demonstrated by our study, subtle, complex relationships within/between clinical data and plan parameters may influence successful treatment delivery. AI and ML have the potential to identify and analyze these relationships for peer review.[23] Similar to the goals of many AI-driven studies, ours is not to replace humans with AI but instead to offer providers additional tools to enhance the process. These can supplement clinical intuition and deepen our understanding of factors that previously might go unanticipated.

Since time is a limited resource, AI/ML can also help to identify and prioritize cases that require more time for discussion. Ultimately, these technologies could improve patient safety and treatment outcomes.

Our study offers a concept that can be used to better identify charts requiring more in-depth review. Many software tools for peer review or chart checking take "rule-based" approaches, meaning that the parameter being flagged will need to be predefined with a range or value.[24] Azmandian et al used clustering techniques for outlier detection based on treatment parameters in four-field box prostate plans; this study helped to detect plan abnormalities without using the rule-based approach.[25] Kalet et al developed a Bayesian network model using clinical and plan parameters to detect errors in radiation therapy plans. Their model utilized a clinical layer (eg, morphology or tumor type), a prescription layer (eg, total dose, dose per fraction, technique), and a treatment layer (eg, monitor unit per fraction, number of beams, beam energy). Their study achieved an AUC of 0.88, 0.98, and 0.89 for the lung, brain, and breast cancer error detection networks, respectively.[26] Similar to our study, Kalet et al found the highest testing set accuracy in the brain subset.

Luk et al also used a Bayesian model incorporating prescription, plan, setup, and diagnostic parameters to detect chart review

errors. The AUC for this study ranged from 0.82 to 0.89.[27] Our testing set accuracy is lower than these other studies, possibly because our cohort included not only "abnormal" cases from a chart-checking perspective but also cases with toxicity-related interruptions.

Although the sensitivity and specificity of our models are relatively low, there are multiple ways in which we can improve these metrics of our study. First, while our study specified remaining fractions as a surrogate for cases likely to experience treatment interruption or complications, remaining fractions can result from reasons unrelated to treatment. Our definition of treatment interruptions as remaining fractions includes patients whose treatments were re-planned and those who discontinued treatment. Separating these cases into two cohorts could improve model predictability.

Second, we grouped subsets based on treatment site, but further dividing them into more focused groups may improve model predictability, eg, separating pelvic plans based on whether they include pelvic lymph nodes or separating spine plans based on vertebral level. Overall, model performance and rigor are expected to improve with increased curation of the dataset and additional clinical factors, including dosimetric plan and structure set data. Additionally, introducing socioeconomic and personal factors, which are commonly seen as causes of treatment interruptions, could improve our models.

Future study directions can also include optimizing hyperparameters and layer structures for the NN. Alternative techniques for data re-balancing or using convolutional NNs for deep learning could also be investigated. A weakness of this study is that it does not include brachytherapy, electron therapy, other treatment sites, or pediatric populations. To address this issue, we are continuing the study with brachytherapy, electrons, and pediatric populations, as well as including additional, more robust parameters.

A known limitation of NNs is their difficulty in identifying the single feature that contributes most to the model, given the complexity of their relationships and the numerous weights/biases assigned to them. Future studies can explore a technique to uncover those features selected by the NN deemed to be most important to the predictive task. At the current stage of our model, which demonstrates moderate predictive performance, making conclusions about or connecting a specific result and/or feature to a reason for a predicted interruption is difficult. Improvements in model performance should enable exploration of more advanced, explainable-AI techniques.

Another current focus of AI and ML is on treatment planning.[23] For example, AI is being used to adapt treatment plans in real time to match the day-to-day variations in patient anatomy, thereby reducing interruptions caused by re-simulation and/or discussion manual re-planning.[28] As the technology for adaptive planning becomes more widely available, our model will likely need additional training to keep up with technological advancements. Future studies may focus on sites where adaptive planning is often beneficial or necessary, such as the head and neck owing to tumor progression or treatment response. The ability to predict cases requiring adaptive planning due to tumor change can allow the care team to anticipate and minimize treatment breaks.

CONCLUSIONS

Our study demonstrated the ability of AI and ML models to predict major changes in patient treatment, including re-planning and radiation therapy cessation. The findings point to the promising capability of AI and ML to augment peer review and encourage further studies in this aspect of radiation oncology.

References

1) Marks LB, Adams RD, Pawlicki T, et al. Enhancing the role of case-oriented peer review to improve quality and safety in radiation oncology: executive summary. *Pract Radiat Oncol.* 2013;3(3):149-156. doi:10.1016/j.prro.2012.11.010

2) Martin-Garcia E, Celada-Álvarez F, Pérez-Calatayud MJ, et al. 100% peer review in radiation oncology: is it feasible? *Clin Transl Oncol.* 2020;22(12):2341-2349. doi:10.1007/s12094-020-02394-8

3) Hoopes DJ, Johnstone PA, Chapin PS, et al. Practice patterns for peer review in radiation oncology. *Pract Radiat Oncol.* 2015;5(1):32-38. doi:10.1016/j.prro.2014.04.004

4) Duggar WN, Bhandari R, Yang CC, Vijayakumar S. Group consensus peer review in radiation oncology: commitment to quality. *Radiat Oncol.* 2018;13(1):55. doi:10.1186/s13014-018-1006-1

5) Shaikh T, Handorf EA, Murphy CT, et al. The impact of radiation treatment time on survival in patients with head and neck cancer. *Int J Radiat Oncol Biol Phys.* 2016;96(5):967-975. doi:10.1016/j.ijrobp.2016.08.046

6) Dahlke S, Steinmann D, Christiansen H, et al. Impact of time factors on outcome in patients with head and neck cancer treated with definitive radio(chemo)therapy. *In Vivo.* 2017;31(5):949-955. doi:10.21873/invivo.11152

7) Lanciano RM, Pajak TF, Martz K, Hanks GE. The influence of treatment time on outcome for squamous cell cancer of the uterine cervix treated with radiation: a patterns-of-care study. *Int J Radiat Oncol.* 1993;25(3):391-397. doi:10.1016/0360-3016(93)90058-4

8) Thaker NG, Sturdevant L, Jhingran A, et al. Assessing the quality of a radiation oncology case-based, peer-review program in an integrated academic and community cancer center network. *J Oncol Pract*. 2016;12(4):e476-86. doi:10.1200/JOP.2015.005983

9) Vijayakumar S, Duggar WN, Packianathan S, Morris B, Yang CC. Chasing zero harm in radiation oncology: using pre-treatment peer review. *Front Oncol*. 2019;9:302. doi:10.3389/fonc.2019.00302

10) Caissie A, Rouette J, Jugpal P, et al. A pan-canadian survey of peer review practices in radiation oncology. *Pract Radiat Oncol*. 2016;6(5):342-351. doi:10.1016/j.prro.2016.01.014

11) Hong JC, Eclov NCW, Dalal NH, et al. System for high-intensity evaluation during radiation therapy (SHIELD-RT): a prospective randomized study of machine learning-directed clinical evaluations during radiation and chemoradiation. *J Clin Oncol*. 2020;38(31):3652-3661. doi:10.1200/JCO.20.01688

12) He HB, Bai Y, Garcia EA, Li ST. *ADASYN: adaptive synthetic sampling approach for imbalanced learning*. IEEE IJCNN; 2008:1322-1328. doi:10.1109/IJCNN.2008.4633969

13) Sher DJ, Posner MR, Tishler RB, et al. Relationship between radiation treatment time and overall survival after induction chemotherapy for locally advanced head-and-neck carcinoma: a subset analysis of tax 324. *Int J Radiat Oncol*. 2011;81(5):e813-e818. doi:10.1016/j.ijrobp.2010.12.005

14) Isaksson LJ, Pepa M, Zaffaroni M, et al. Machine learning-based models for prediction of toxicity outcomes in radiotherapy. *Front Oncol*. 2020;10:790. doi:10.3389/fonc.2020.00790

15) Zhang B, Shi H, Wang H. Machine learning and AI in cancer prognosis, prediction, and treatment selection: a critical approach. *J Multidiscip Healthc*. 2023;16:1779-1791. doi:10.2147/JMDH.S410301

16) Reddy J, Lindsay WD, Berlind CG, Ahern CA, Smith BD. Applying a machine learning approach to predict acute toxicities during radiation for breast cancer patients. *Int J Radiat Oncol*. 2018;102(3):S59. doi:10.1016/j.ijrobp.2018.06.167

17) Das SK, Chen SF, Deasy JO, et al. Combining multiple models to generate consensus: application to radiation-induced pneumonitis prediction. *Med Phys*. 2008;35(11):5098-5109. doi:10.1118/1.2996012

18) Reddy JP, Lindsay WD, Berlind CG, et al. Applying a machine learning approach to predict acute radiation toxicities for head and neck cancer patients (vol 105, pg S69, 2019). *Int J Radiat Oncol Jan*. 106(1):223-223. doi:10.1016/j.ijrobp.2019.10.013

19) Jiang W, Lakshminarayanan P, Hui X, et al. Machine learning methods uncover radiomorphologic dose patterns in salivary glands that predict xerostomia in patients with head and neck cancer. *Adv Radiat Oncol*. 2019;4(2):401-412. doi:10.1016/j.adro.2018.11.008

20) Carrara M, Massari E, Cicchetti A, et al. Development of a ready-to-use graphical tool based on artificial neural network classification: application for the prediction of late fecal incontinence after prostate cancer radiation therapy. *Int J Radiat Oncol*. 2018;102(5):1533-1542. doi:10.1016/j.ijrobp.2018.07.2014

21) Pella A, Cambria R, Riboldi M, et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. *Med Phys*. 2011;38(6):2859-2867. doi:10.1118/1.3582947

22) Lawrence YR, Whiton MA, Symon Z, et al. Quality assurance peer review chart rounds in 2011: a survey of academic institutions in the united states. *Int J Radiat Oncol Biol Phys*. 2012;84(3):590-595. doi:10.1016/j.ijrobp.2012.01.029

23) Kiser KJ, Fuller CD, Reed VK. Artificial intelligence in radiation oncology treatment planning: a brief overview. *J Med Artif Intell*. 2019;2:9-9. doi:10.21037/jmai.2019.04.02

24) Luk SMH, Ford EC, Phillips MH, Kalet AM. Improving the quality of care in radiation oncology using artificial intelligence. *Clin Oncol*. 2022;34(2):89-98. doi:10.1016/j.clon.2021.11.011

25) Azmandian F, Kaeli D, Dy JG, et al. Towards the development of an error checker for radiotherapy treatment plans: a preliminary study. *Phys Med Biol*. 2007;52(21):6511-6524. doi:10.1088/0031-9155/52/21/012

26) Kalet AM, Gennari JH, Ford EC, Phillips MH. Bayesian network models for error detection in radiotherapy plans. *Phys Med Biol*. 2015;60(7):2735-2749. doi:10.1088/0031-9155/60/7/2735

27) Luk SMH, Meyer J, Young LA, et al. Characterization of a bayesian network-based radiotherapy plan verification model. *Med Phys*. 2019;46(5):2006-2014. doi:10.1002/mp.13515

28) Wu QJ, Li T, Wu Q, Yin FF. Adaptive radiation therapy: technical components and clinical applications. *Cancer J*. 2011;17(3):182-189. doi:10.1097/PPO.0b013e31821da9d8