MACHINE LEARNING WITH RADIATION ONCOLOGY BIG DATA

EDITED BY: Jun Deng, Issam El Naqa and Lei Xing
PUBLISHED IN: Frontiers in Oncology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-730-4
DOI 10.3389/978-2-88945-730-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

MACHINE LEARNING WITH RADIATION ONCOLOGY BIG DATA

Topic Editors:
Jun Deng, Yale University, United States
Issam El Naqa, University of Michigan, United States
Lei Xing, Stanford University, United States

Image: Immersion Imagery/Shutterstock.com

Citation: Deng, J., Naqa, I. E., Xing, L., eds. (2019). Machine Learning With Radiation Oncology Big Data. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-730-4

Table of Contents

frontiers in Oncology

EDITORIAL
published: 27 September 2018
doi: 10.3389/fonc.2018.00416

Editorial: Machine Learning With Radiation Oncology Big Data

Jun Deng[1]*, *Issam El Naqa*[2] *and Lei Xing*[3]

[1] Department of Therapeutic Radiology, Yale University, New Haven, CT, United States, [2] Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, United States, [3] Department of Radiation Oncology, Stanford University, Stanford, CA, United States

Keywords: big data, machine learning, artificial intelligence, personalized medicine, personalized radiotherapy

Editorial on the Research Topic

Machine Learning With Radiation Oncology Big Data

INTRODUCTION

Half of all cancer patients may receive radiotherapy as part of their treatment. With the wealth of diverse data generated every day in the clinic, the radiation oncology community possesses a unique advantage in harnessing these massive data with the predictive power of machine learning methods for the benefit of millions of cancer patients undergoing radiotherapy worldwide. In this Research Topic "Machine Learning with Radiation Oncology Big Data," a wide range of clinical applications involving various machine learning algorithms have been described and demonstrated, with the hope of ushering in more widespread applications of artificial intelligence in medicine, particularly in cancer radiotherapy in order to achieve a truly individualized radiation oncology and an evidence-based learning healthcare system.

TOPICS COVERED IN THIS RESEARCH TOPIC

- Knowledge-based treatment planning: Zhang et al.
- Knowledge-based response-adapted radiotherapy: Tseng et al.
- Radiomics image analysis: Elhalawani et al., Jethanandani et al.
- Radiogenomics and outcome modeling: Kang et al.
- Automated contouring and nodule detection: Jackson et al., Ali et al., Men et al.
- Comparative effectiveness research: Sanders and Showalter.
- Machine learning in radiation oncology overview: Feng et al.
- Normal tissue complication probability modeling: Gabryś et al.

PAPERS INCLUDED IN THIS RESEARCH TOPIC

In this review paper, Elhalawani et al. summarized the feedback of eight contestants who participated in a recent radiomics challenge in head and neck radiation oncology, and discussed some of the challenges in sharing and directing existing datasets toward clinical implementation of radiomics in radiation oncology.

Tseng et al. discussed recent development in the knowledge-based response-adapted radiotherapy for personalized radiotherapy management. They addressed three specific questions that are necessary to realize it clinically: (1) what knowledge is needed, (2) how to estimate radiotherapy outcomes accurately, and (3) how to adapt optimally.

Kang et al. presented an overview of machine learning algorithms in the application of radiogenomics to combine genomics signatures with radiotherapy. They summarized

OPEN ACCESS

Edited and reviewed by:
Timothy James Kinsella, Warren Alpert Medical School of Brown University, United States

***Correspondence:**
Jun Deng
jun.deng@yale.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *04 September 2018*
Accepted: *07 September 2018*
Published: *27 September 2018*

Citation:
Deng J, El Naqa I and Xing L (2018) Editorial: Machine Learning With Radiation Oncology Big Data. Front. Oncol. 8:416. doi: 10.3389/fonc.2018.00416

the important lessons learned for the proper integration of machine learning into radiogenomics analysis.

Jackson et al. introduced a convolutional neural network approach for fully automated contouring of kidneys and automated radiation dose estimation in an unsealed source therapy, which provides comparable accuracy to humans while largely reducing the planning time.

In a systematic review, Jethanandani et al. explored the various applications of radiomics in magnetic resonance imaging of head and neck cancer, and identified the lack of standardization in study design as a major limitation to their clinical relevance.

Sanders and Showalter described their vision of combining big data with comparative effectiveness research methodologies within the framework of a rapid-learning healthcare system in order to accelerate discovery and realize a fully individualized radiation treatment.

Feng et al. identified specific opportunities in a long chain of radiotherapy processes where machine learning could improve the quality and efficiency of patient care in radiation oncology, as well as the needs required to realize them at both the community and institutional levels.

Ali et al. presented a robust non-invasive deep reinforcement learning method to predict the presence of lung nodules, a common precursor to lung cancer, based on 888 lung CT scans of the lung nodule analysis (LUNA) challenge.

Zhang et al. proposed an ensemble approach to knowledge-based intensity modulated radiation therapy treatment planning, and demonstrated its advantages in terms of robustness against small training set sizes, mis-labeled cases, and dosimetric inferior plans.

Gabryś et al. investigated whether machine learning with dosimetric, radiomic, and demographic features can allow for more precise xerostomia risk assessment. They identified the need for the development of personalized data-driven risk profiles for normal tissue complication probability (NTCP) modeling.

Men et al. developed an end-to-end deep deconvolutional neural network for segmentation of nasopharyngeal tumor volumes to improve the consistency of contouring and streamline radiotherapy workflows, but cautioned that careful human review and a considerable amount of editing would still be required.

CONCLUSIONS AND OUTLOOK

The 11 papers included in this Research Topic produced some promising results and offered visionary perspectives regarding the role of machine learning with radiation oncology big data. The clinical applications demonstrated here are considered just the tip of the iceberg of the incoming full-spectrum applications of human intelligence and artificial intelligence in radiation oncology. While still in its infancy stage, we envisage that artificial intelligence together with human intelligence can provide something much better than either one could perform alone in the near future.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 20 December 2017
doi: 10.3389/fonc.2017.00315

Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images

Kuo Men, Xinyuan Chen, Ye Zhang, Tao Zhang, Jianrong Dai, Junlin Yi* and Yexiong Li*

National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Background: Radiotherapy is one of the main treatment methods for nasopharyngeal carcinoma (NPC). It requires exact delineation of the nasopharynx gross tumor volume (GTVnx), the metastatic lymph node gross tumor volume (GTVnd), the clinical target volume (CTV), and organs at risk in the planning computed tomography images. However, this task is time-consuming and operator dependent. In the present study, we developed an end-to-end deep deconvolutional neural network (DDNN) for segmentation of these targets.

Methods: The proposed DDNN is an end-to-end architecture enabling fast training and testing. It consists of two important components: an encoder network and a decoder network. The encoder network was used to extract the visual features of a medical image and the decoder network was used to recover the original resolution by deploying deconvolution. A total of 230 patients diagnosed with NPC stage I or stage II were included in this study. Data from 184 patients were chosen randomly as a training set to adjust the parameters of DDNN, and the remaining 46 patients were the test set to assess the performance of the model. The Dice similarity coefficient (DSC) was used to quantify the segmentation results of the GTVnx, GTVnd, and CTV. In addition, the performance of DDNN was compared with the VGG-16 model.

Results: The proposed DDNN method outperformed the VGG-16 in all the segmentation. The mean DSC values of DDNN were 80.9% for GTVnx, 62.3% for the GTVnd, and 82.6% for CTV, whereas VGG-16 obtained 72.3, 33.7, and 73.7% for the DSC values, respectively.

Conclusion: DDNN can be used to segment the GTVnx and CTV accurately. The accuracy for the GTVnd segmentation was relatively low due to the considerable differences in its shape, volume, and location among patients. The accuracy is expected to increase with more training data and combination of MR images. In conclusion, DDNN has the potential to improve the consistency of contouring and streamline radiotherapy workflows, but careful human review and a considerable amount of editing will be required.

Keywords: automatic segmentation, target volume, deep learning, deep deconvolutional neural network, radiotherapy

OPEN ACCESS

Edited by:
Jun Deng,
Yale University,
United States

Reviewed by:
Wenyin Shi,
Thomas Jefferson University,
United States
Marianne Aznar,
University of Manchester,
United Kingdom

****Correspondence:***
Jianrong Dai
dai_jianrong@163.com;
Junlin Yi
yijunlin1969@163.com

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *25 August 2017*
Accepted: *05 December 2017*
Published: *20 December 2017*

Citation:
Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J and Li Y (2017) Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. Front. Oncol. 7:315. doi: 10.3389/fonc.2017.00315

INTRODUCTION

Nasopharyngeal carcinoma (NPC) is a malignant tumor prevalent in southern China. Radiotherapy is one of the main treatments for NPC, and its rapid development has played a significant role in the improvement of tumor control probability. Intensity-modulated radiotherapy and volumetric-modulated radiotherapy (VMAT) have become the state-of-the-art methods for the treatment of NPC over the past two decades (1, 2). These technologies can facilitate dose escalation to the tumor target while improving the sparing of organs at risk (OARs), and the dose distribution usually has steep gradients at the target boundary. Modern treatment planning system (TPS) requires exact delineation of the nasopharynx gross tumor volume (GTVnx), the metastatic lymph node gross tumor volume (GTVnd), the clinical target volume (CTV) to be irradiated, and OARs to be spared in planning computed tomography (CT) images so that a radiation delivery plan can be optimized reversely. This task is a type of image segmentation and is usually carried out manually by radiation oncologists based on recommended guidelines (e.g., RTOG 0615 Protocol). However, the manual segmentation (MS) process is time-consuming and operator dependent. It has been reported that the segmentation of a single head-and-neck (H&N) cancer case takes an average of ~2.7 h (3). This time-consuming work may be repeated several times during a course of NPC radiotherapy due to a tumor response or significant anatomic changes and alterations. In addition, the accuracy of the segmentation is highly dependent on the knowledge, experience, and preference of the radiation oncologists. Considerable inter- and intra-observer variation in segmentation of these regions of interest (ROIs) have been noted in a number of studies (4–7).

As a result, a fully automated segmentation method for radiotherapy is helpful to relieve radiation oncologists from the labor-intensive aspects of their work and increase the accuracy, consistency, and reproducibility of ROI delineation. "Atlas-based segmentation" (ABS) (8–10) incorporates a prior knowledge into the process of segmentation and is one of the most widely used and successful image segmentation techniques for biomedical applications. In this type of method, an optimal transformation between the target image to be segmented and a single atlas or multiple atlases containing some ground truth segmentations is computed using deformable registration techniques. Then, all the labeled structures in the atlas image can be propagated through the registration transformation onto the target image automatically. ABS has become a popular method in automatic delineation of target and/or OARs in H&N radiotherapy (11–17) due to its acceptable results and fully unsupervised mode of operation. Han et al. (11) used the object shape information in the atlas to account for large inter-subject shape differences. Sjöberg et al. (12) applied fusion of multiple atlases to improve the segmentation accuracy than single atlas segmentation. Tao et al. (13) used ABS to reduce interobserver variation and improve dosimetric parameter consistency for OARs. Teguh et al. (14) evaluated autocontouring using ABS and found it was a useful tool for rapid delineation, although editing was inevitable. Sims et al. (15) did a pre-clinical assessment of ABS and showed that it exhibited satisfactory sensitivity; however, careful review and editing were required. Walker et al. (16) concluded ABS was timesaving in generating ROI in H&N, but attending physician approval remained vital. However, there are two main challenges using the ABS method. First, due to the anatomical variations of human organs, it is difficult to build a "universal atlas" for all human organs. The ROI may be considerably different according to the body shape and body size of the patient. The variability should be taken into account to construct a patient-specific atlas from all atlas images, but there are difficulties for target images with a large variability in shape and appearance. Second, a large disadvantage of using ABS is the large computation time that is involved in registering the target image to its atlas image (18). Moreover, it often requires the target image to be aligned to multiple atlases, which will increase the process of registration several times.

Deep learning methods have achieved enormous success in many computer vision tasks, such as image classification (19–21), object detection (22, 23), and semantic segmentation (24–26). Convolutional neural networks (CNNs) have become the most popular algorithm for deep learning (21, 27). CNNs consist of alternating convolutional and pooling layers to automatically extract multiple-level visual features and have made significant progress in computer-aided diagnosis and automated medical image analysis (28–31). Melendez et al. (29) applied multiple-instance learning for tuberculosis detection using chest X-rays and reported an AUC of 0.86. Hu et al. (30) proposed a liver segmentation framework based on CNNs and globally optimized surface evolution, yielding a mean Dice similarity coefficient (DSC) of 97%. Esteva et al. (31) trained a CNN using a large dataset to classify skin cancer and achieved higher accuracy than dermatologists. In addition, CNNs have been applied in the segmentation of many organs and substructures, such as cells (32), nuclei (33), blood vessels (34), neuronal structures (35), brain (36), ventricles (37), liver (38), kidneys (39), pancreas (40), prostate gland (41), bladder (42), colon (43), and vertebrae (44) with relatively better overlap compared with state-of-the-art methods. However, these studies have been confined mostly to the field of radiology.

Furthermore, there has been increasingly more interest in applying CNNs to radiation therapy (45–48). Recently, Ibragimov and Xing (49) used CNNs for OARs segmentation in H&N CT images and obtained DSC values that varied from 37.4% for chiasm to 89.5% for mandible. This was the first report on OAR delineation with CNNs in radiotherapy; however, no target was segmented. In this work, we developed a deep deconvolutional neural network (DDNN) for the segmentation of CTV, GTVnx, and GTVnd for radiotherapy of NPC. The experimental results show that the DDNN can be used to realize the segmentation of NPC targets while planning CT images. DDNN is an end-to-end

Abbreviations: NPC, nasopharyngeal carcinoma; GTVnx, nasopharynx gross tumor volume; GTVnd, metastatic lymph node gross tumor volume; CTV, clinical target volume; OARs, organs at risk; CT, computed tomography; H&N, head and neck; GT, ground truth; CNNs, convolutional neural networks; DDNN, deep deconvolutional neural network; DSC, Dice similarity coefficient; TCP, tumor control probability; IMRT, intensity-modulated radiotherapy; VMAT, volumetric-modulated radiotherapy; TPS, treatment planning system; ROIs, regions of interest; ABS, Atlas-based segmentation; BN, batch normalized; ReLU, rectified linear non-linearity.

architecture consisting of two important components, including an encoder and a decoder. Different from typical CNNs, we performed a reversed deconvolution at decoder networks to rebuild high-resolution feature maps from low-resolution ones. Our work is the first attempt at applying DDNN for the auto-segmentation of a target for the planning of radiotherapy in NPC.

MATERIALS AND METHODS

Data Acquisition

A total of 230 patients diagnosed with NPC stage I or stage II that received radiotherapy during January 2011 to January 2017 in our department were included in our study. All patients were immobilized with a thermoplastic mask (head, neck, shoulder) in the supine position. Simulation contrast CT data were acquired on a Somatom Definition AS 40 (Siemens Healthcare, Forchheim, Germany) or Brilliance CT Big Bore (Philips Healthcare, Best, the Netherlands) system set on helical scan mode with contrast enhancement. CT images were reconstructed using a matrix size of 512 × 512 and thickness of 3.0 mm. MR images of all patients were acquired to assist the definition of the targets. Radiation oncologists contoured the GTVnx, the GTVnd, CTV, and OARs in the planning CT using a Pinnacle TPS (Philips Radiation Oncology Systems, Fitchburg, WI, USA) system. The GTVnx was defined as the primary nasopharyngeal tumor mass. The GTVnd was defined as the metastatic lymph nodes. The CTV (CTV1 + CTV2) included GTVnx, GTVnd, high-risk local regions that contain the parapharyngeal spaces, the posterior third of nasal cavities and maxillary sinuses, pterygoid processes, pterygopalatine fossa, the posterior half of the ethmoid sinus, cavernous sinus, base of skull, sphenoid sinus, the anterior half of the clivus, petrous tips, and high-risk lymphatic drainage areas, including bilateral retropharyngeal lymph nodes and level II.

DDNN Model for Segmentation

In the present study, we introduced a DDNN model to segment the target NPC for radiotherapy. DDNN is an end-to-end segmentation framework that can predict pixel class labels in CT images. **Figure 1** depicts the flowchart of the proposed model. As is shown in **Figure 2**, the DDNN networks consisted of two important components, including an encoder part and a decoder part. The encoder network consisted of 13 convolutional layers for feature extraction and was used to extract the visual features of the medical image, and the decoder network recovered the original resolution by deploying deconvolution. Specifically, the

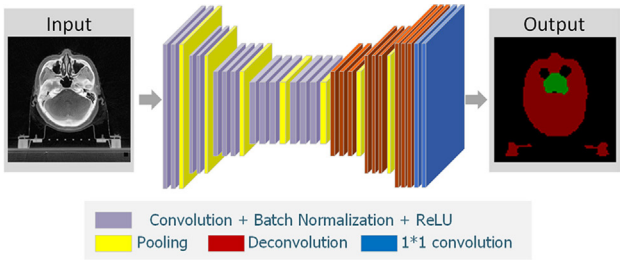

FIGURE 1 | Overall framework of the proposed algorithm.

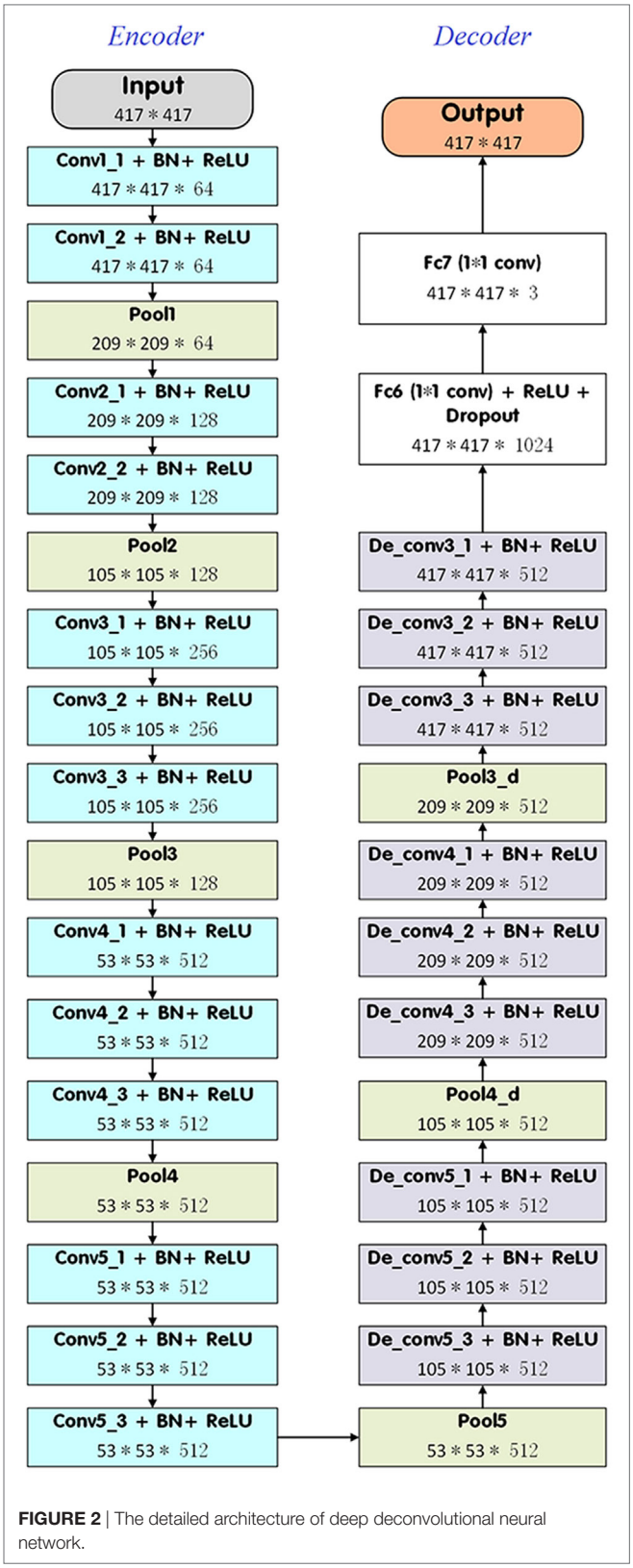

FIGURE 2 | The detailed architecture of deep deconvolutional neural network.

encoder network layers were based on the VGG-16 architecture (21), used for high-quality image classification. Different from VGG-16, we performed a reversed deconvolution at decoder networks to rebuild high-resolution feature maps from low-resolution. In addition, we replaced the fully connected layers with fully convolutional layers for our segmentation task. With the adaptation, the networks can achieve pixel segmentation in CT images. Please refer to the appendix for more technical specifications of the architecture.

Experiments

Data from 184 patients out of 230 were chosen randomly as a training set to adjust the parameters of the DDNN model, and the remaining 46 patients were used as the test set to evaluate the performance of the model. In this work, we implemented our model's training, evaluation, error analysis, and visualization pipeline using Caffe (50), which is a popular deep learning framework, and then compiled using cuDNN (51) computational kernels. For the experiments, we adopted data augmentation techniques, such as random cropping and flipping to reduce over fitting. We used stochastic gradient descent with momentum to optimize the loss of function. We set the initial learning rate to 0.0001, learning rate decay factor to 0.0005, and decay step size to 2,000. Instead of using a fixed number of steps, we trained our model until the mean average precision of the training set converged, and then evaluated the model using the validation set. We used NVIDIA TITAN XP GPU for all experiments.

Quantitative Evaluation

A total of 46 patients were used to assess the performance of the model. MSs were defined as the reference segmentations generated by the experienced radiation oncologists. All the voxels that belong to the MS were extracted and labeled. During the testing phase, all the 2D CT slices were tested one by one. The input was the 2D CT image, and the final output was pixel-level classification, which was the most likely classification label. Performance of the proposed method was tested and compared with the segmentation of the GTVnx, GTVnd, and CTV. The DSC and the Hausdorff distance (H) were used to quantify the results.

The DSC is defined as shown in Eq. 1 as follows:

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

where A represents the MS, B denotes the auto-segmented structure and A ∩ B is the intersection of A and B. The DSC results in values between 0 and 1, where 0 represents no intersection at all and 1 reflects perfect overlap of structures A and B.

The Hausdorff distance (H) is defined as

$$H(A,B) = \max(h(A,B), h(B,A)) \quad (2)$$

where

$$h(A,B) = \max(\min_{a \in A, b \in B} \| a - b \|) \quad (3)$$

and ‖.‖ is some underlying norm on the points of A and B. As H(A,B) diminishes, the overlap between A and B increases.

In addition, the performance of DDNN was compared with VGG-16. The average DSC and Hausdorff distance values for the three targets (GTVnx, GTVnd, and CTV) were analyzed with paired t-tests between DDNN and VGG-16. All analyses were performed with a p-value set to <0.05.

RESULTS

The results for all tested patients and GTVnx, GTVnd, and CTV values are summarized in **Figure 3** and **Table 1**. The proposed DDNN auto-segmentation showed a better overall agreement than the VGG-16 based auto-segmentation, as shown by the DSC values. The average DSC value of DDNN was 15.4% higher than the VGG-16 average DSC value (75.3 ± 11.3 vs. 59.9 ± 22.7%, $p < 0.05$). Automatic delineation with DDNN produced a good result for the GTVnx and CTV, with DSC values of 80.9 and 82.6%, respectively. These values showed a reasonable volume overlap of the auto-segmented contours and the manual contours. The quality of the automatically generated GTVnd was barely satisfactory, with a mean DSC value of 62.3%. The Hausdorff distance values for all targets were reduced by DDNN compared with VGG-16 (12.6 ± 11.5 vs. 23.4 ± 24.4, $p < 0.05$).

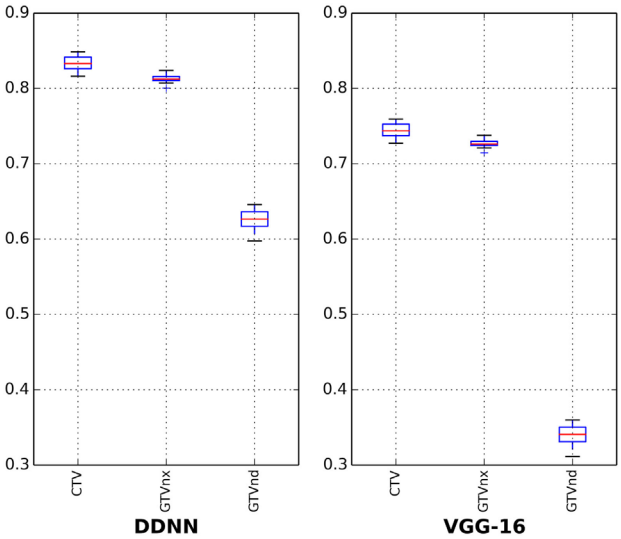

FIGURE 3 | Boxplots obtained from Dice similarity coefficient analyses.

TABLE 1 | Dice similarity coefficient (DSC) and Hausdorff distance for nasopharynx gross tumor volume (GTVnx), metastatic lymph node gross tumor volume (GTVnd), and clinical target volume (CTV).

	DSC (%)			Hausdorff distance (mm)		
Region of interest	**CTV**	**GTVnx**	**GTVnd**	**CTV**	**GTVnx**	**GTVnd**
Deep deconvolutional neural network	82.6	80.9	62.3	6.9	5.1	25.8
VGG-16	73.7	72.3	33.7	11.1	7.7	51.5

Figures 4–6 show auto-segmentation of CTV, GTVnx, and GTVnd for test cases, respectively. In these examples, the auto-segmented contours of CTV and GTVnx using DDNN were close to the MS contours, although inconsistencies existed. Only a few corrections were necessary to validate the automatic segmentation. However, for the segmentation of the GTVnd, there was some deviation from the MS in shape, volume, and location.

DISCUSSION

We have designed an automated method to segment CT images of NPC. To the best of our knowledge, this task has not previously been reported. Our results suggest that the proposed DDNN algorithm can learn the semantic information from nasopharyngeal CT data and produce high-quality segmentation of the target. We compared the proposed architecture with the popular Deeplab v2 VGG-16 model. This comparison revealed that our method achieved better segmentation performance. Our DDNN method deployed a deeper encoder and decoder neural network, which used convolutional filters to extract feature and deployed deconvolutional filters to recover the original resolution. Thus, detailed segmented results were learned/predicted better than bilinear interpolation.

Consistency of target delineation is essential for the improvement of radiotherapy outcomes. Leunens et al. (52) demonstrated that inter- and intra-observer variation is considerable. Lu et al. (53) investigated the interobserver variations in GTV contouring of H&N patients and reported a DSC value of only 75%. Caravatta et al. (54) evaluated the overlap accuracy of CTV delineation among different radiation oncologists and got a DSC of 68%. Automatic segmentation has the potential to reduce variability of contours among physicians and improve efficiency. The gains in efficiency and consistency are valuable only if accuracy is not compromised. Assessment of accuracy of a segmentation method is complex, because there is no common database or objective volume for comparison. The evaluation of automatic segmentation for radiotherapy planning usually uses the DSC value, thus providing a reasonable basis for comparison. Apparently, our method showed good performance compared with the existing studies regarding the auto-segmentation topic. In addition, such auto-segmentation methods are atlas- and/or model based, and there is no report on segmentation of GTV or CTV using a deep learning method. Regarding the target, the comparison is difficult since *N*-stage (most often *N*0) and selected levels were quite different from one study to another. For CTV, different previous publications reported mean DSC values of 60% (55), 60% (8), 60% (56), 67% (14), 77% (57), 78% (58), 79% (59), and 80.2% (60), whereas the DSC value of DDNN was 82.6%. There are few reports on auto-segmentation of GTVnx or GTVnd. For segmentation of GTVnx, DSC values

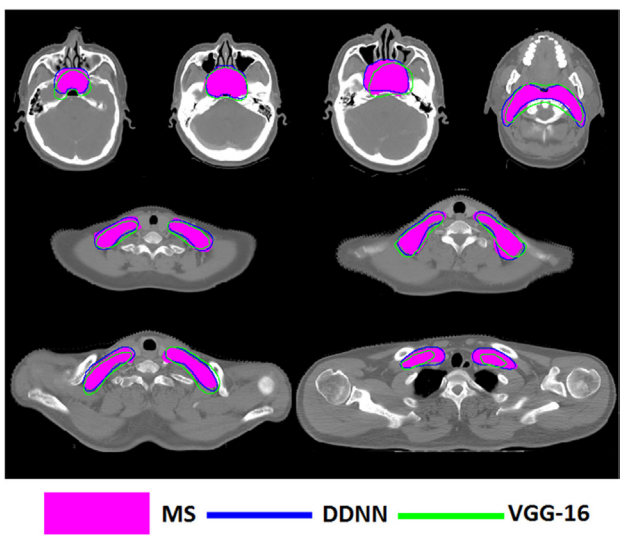

FIGURE 4 | Segmentation results for clinical target volume, shown in transverse computed tomography slices.

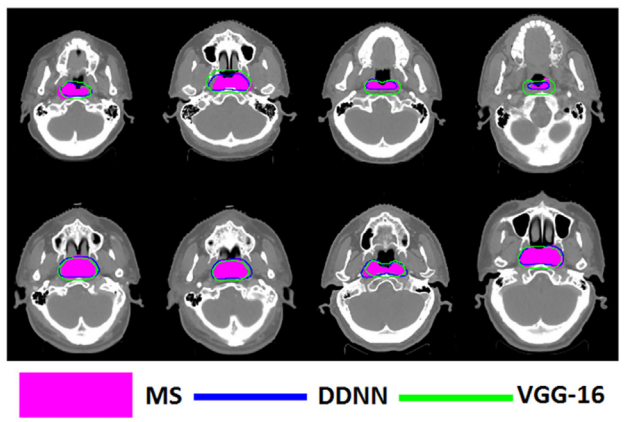

FIGURE 5 | Segmentation results for nasopharynx gross tumor volume, shown in transverse computed tomography slices.

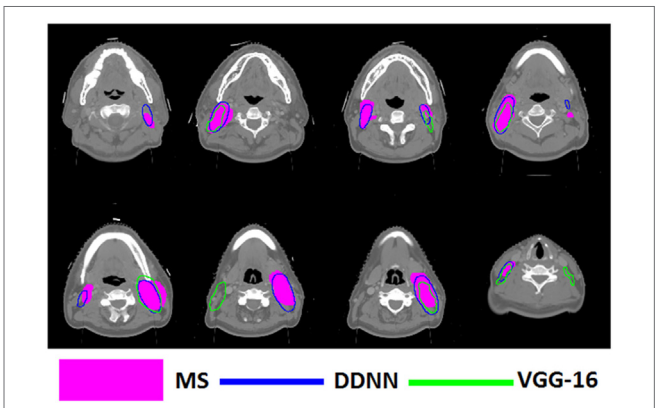

FIGURE 6 | Segmentation results for metastatic lymph node gross tumor volume, shown in transverse computed tomography slices.

have been reported to be 69.0% (58) and 75.0% (61), whereas our proposed method demonstrated a high DSC value of 80.9%. The segmentation of GTVnd reported in the literature has yielded DSC values of 46.0% (62), and our method showed a DSC value of 62.3%. It is unfair to say our proposed algorithm is superior because the comparison with the published methods was not done with the same dataset; however, it is reasonable to conclude DDNN resulted in good results. Meanwhile, the proposed method learns and predicts in an end-to-end form without post-processing, which makes the inference time of the whole network within seconds.

Although the segmentation accuracy for GTVnd was better than previously reported, it was still too low. There are several reasons for this deficiency. First, this low result was due to lack of soft tissue contrast in CT-based delineation. Second, the GTVnd typically does not have constant image intensity or clear anatomic boundaries, and its shape and location are more variable compared with CTV and GTVnx among different patients. Moreover, there is no GTVnd region in *N*0 patients, who were also included in our training and test sets. All of these factors will hinder the DDNN model from learning the robust features and making accurate reasoning. Thus, the segmentation accuracy of GTVnd remains unsatisfactory at present. Zijdenbos et al. (63) suggests that a DSC value of >70% represents good overlap. Although the segmentation accuracy of CTV and GTVnx exceeded this standard, attending physician oversight remains critical. Imperfect definition of target volumes, which are then used for treatment planning, may result in under dosage of target volumes or an overdose delivered to normal tissues. As a result, the proposed method cannot be applied in an unsupervised fashion in the clinic. Human review and a considerable amount of editing might be required.

There are several limitations to our study. First, a model trained on *N*0 and *N*+ patients was used to assess the testing set, including both *N*0 and *N*+ patients. This may make the model difficult to converge and reduce the accuracy of the prediction. Second, only one physician delineated the target for each patient but all the patients were delineated by several observers. Although the targets were contoured by experts according to the same guideline for NPC, there was still interobserver variability in all cases. We cannot exclude such possible bias, which challenges the DDNN method. Another limitation of our study is that all of the included patients were stage I or stage II. A target with different stages may have different contrast, shapes, and volumes, thus, influencing the performance of the automated segmentation.

This study mainly focused on NPC target segmentation from CT images. However, MR images in H&N have superior soft-tissue contrast and the GTV delineation often depends on MR images. In addition, functional MR may allow accurate location of the tumors. In the future, DDNN is expected to combine with the MR or other types of images to improve target volume delineation. The training set included only 184 patients. Increasing the amount of training data could make the DDNN model more robust, improving the segmentation accuracy. With the initiation of improved target visualization and further improvement of segmentation algorithms in the future, accuracy of auto-segmentation is likely to improve.

CONCLUSION

Accurate and consistent delineation of tumor target and OARs is particularly important in radiotherapy. Several studies have focused on the segmentation of OARs using deep learning methods. This study shows a method using DDNN architecture to auto-segment nasopharyngeal cancer stage I or stage II in planning CT images. The results suggest that DDNN can be used to segment GTVnx and CTV with high accuracy. The accuracy for GTVnd segmentation was relatively low due to the considerable differences in shape, volume, and location among patients. The performance is expected to improve with multimodality medical images and more training data. In conclusion, DDNN has the potential to improve the consistency of contouring and streamline radiotherapy workflows, but careful human review and a considerable amount of editing will be required.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and/or analyzed during the current study are not publicly available due to data security but are available from the corresponding author on reasonable request.

ETHICS STATEMENT

This study was carried out in accordance with the Declaration of Helsinki and was approved by the Independent Ethics Committee of Cancer Hospital, Chinese Academy of Medical Sciences with the following reference number: NCC2015 YQ-15.

AUTHOR CONTRIBUTIONS

All authors discussed and conceived of the study design. KM wrote the programs and performed data analysis, and drafted the manuscript. XC and YZ analyzed and interpreted the patients' data. JD, JY, and YL guided the study and participated in discussions and preparation of the manuscript. All authors read, discussed, and approved the final manuscript.

ACKNOWLEDGMENTS

The authors sincerely thank Dr. Junge Zhang and Dr. Peipei Yang of Institute of Automation, Chinese Academy of Sciences, Dr. Kangwei Liu of LeSee & Faraday Future AI Institute, and Mr. Rongliang Cheng of Qingdao University of Science and Technology for data mining and editing the manuscript. They also thank the radiation oncologists in our department for the target delineation.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 11605291 and No. 11475261), the National Key Projects of Research and Development of China (No. 2016YFC0904600 and No. 2017YFC0107501), and the Beijing Hope Run Special Fund of Cancer Foundation of China (No. LC2015B06).

REFERENCES

1. Lee TF, Ting HM, Chao PJ, Fang FM. Dual arc volumetric-modulated arc radiotherapy (VMAT) of nasopharyngeal carcinomas: a simultaneous integrated boost treatment plan comparison with intensity-modulated radiotherapies and single arc VMAT. *Clin Oncol* (2011) 24(3):196–207. doi:10.1016/j.clon.2011.06.006
2. Moretto F, Rampino M, Munoz F, Ruo Redda MG, Reali A, Balcet V, et al. Conventional 2D (2DRT) and 3D conformal radiotherapy (3DCRT) versus intensity-modulated radiotherapy (IMRT) for nasopharyngeal cancer treatment. *Radiol Med* (2014) 119(8):634–41. doi:10.1007/s11547-013-0359-7
3. Harari PM, Shiyu S, Wolfgang AT. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. *Int J Radiat Oncol Biol Phys* (2010) 77(3):950–8. doi:10.1016/j.ijrobp.2009.09.062
4. Breen S, Publicover J, de Silva S, Pond G, Brock K, O'Sullivan B, et al. Intraobserver and interobserver variability in GTV delineation on FDG-PET-CT images of head and neck cancers. *Int J Radiat Oncol Biol Phys* (2007) 68(3):763–70. doi:10.1016/j.ijrobp.2006.12.039
5. Feng MU, Demiroz C, Vineberg KA, Balter JM, Eisbruch A. Intra-observer variability of organs at risk for head and neck cancer: geometric and dosimetric consequences. *Fuel Energy Abstr* (2010) 78(3):S444–5. doi:10.1016/j.ijrobp.2010.07.1044
6. Yamazaki H, Hiroya S, Takuji T, Naohiro K. Quantitative assessment of inter-observer variability in target volume delineation on stereotactic radiotherapy treatment for pituitary adenoma and meningioma near optic tract. *Radiat Oncol* (2011) 6(1):10. doi:10.1186/1748-717X-6-10
7. Vinod SK, Myo M, Michael GJ, Lois CH. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *J Med Imaging Radiat Oncol* (2016) 60(3):393–406. doi:10.1111/1754-9485.12462
8. Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. *Med Image Anal* (2015) 24(1):205–19. doi:10.1016/j.media.2015.06.012
9. Cabezas M, Oliver A, Lladó X, Freixenet J, Cuadra MB. A review of atlas-based segmentation for magnetic resonance brain images. *Comput Methods Programs Biomed* (2011) 104(3):e158–77. doi:10.1016/j.cmpb.2011.07.015
10. Cuadra MB, Duay V, Thiran JP. *Atlas-Based Segmentation. Handbook of Biomedical Imaging*. USA: Springer (2015). p. 221–44.
11. Han X, Hoogeman MS, Levendag PC, Hibbard LS, Teguh DN, Voet P, et al. Atlas-based auto-segmentation of head and neck CT images. *International Conference on Medical Image Computing and Computer-assisted Intervention*. Berlin, Heidelberg: Springer (2008). p. 434–41.
12. Sjöberg C, Martin L, Christoffer G, Silvia J, Anders A, Anders M. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. *Radiat Oncol* (2013) 8(1):229. doi:10.1186/1748-717X-8-229
13. Tao CJ, Yi JL, Chen NY, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. *Radiother Oncol* (2015) 115(3):407–11. doi:10.1016/j.radonc.2015.05.012
14. Teguh DN, Levendag PC, Voet PW, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *Int J Radiat Oncol Biol Phys* (2010) 81(4):950–7. doi:10.1016/j.ijrobp.2010.07.009
15. Sims R, Isambert A, Grégoire V, Bidault F, Fresco L, Sage J, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. *Radiother Oncol* (2009) 93(3):474–8. doi:10.1016/j.radonc.2009.08.013
16. Walker GV, Awan M, Tao R, Koay EJ, Boehling NS, Grant JD, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. *Radiother Oncol* (2014) 112(3):321–5. doi:10.1016/j.radonc.2014.08.028
17. Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med Phys* (2017) 44(5):2020–36. doi:10.1002/mp.12197
18. Langerak TR, Berendsen FF, Van der Heide UA, Kotte AN, Pluim JP. Multiatlas-based segmentation with preregistration atlas selection. *Med Phys* (2013) 40(9):091701. doi:10.1118/1.4816654
19. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. Nevada, United States (2012). p. 1097–105.
20. Sermanet P, David E, Zhang X, Michael M, Rob F, Yann L. *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks*. (2014) arXiv:1312.6229.
21. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-scale Image Recognition*. (2014) arXiv:1409.1556.
22. Girshick R, Jeff D, Trevor D, Jitendra M. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, United States (2014). p. 580–7.
23. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *In Advances in Neural Information Processing Systems*. Montreal, Canada (2015). p. 91–9.
24. Chen LC, George P, Iasonas K, Kevin M, Alan LY. *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs*. (2014) arXiv:1412.7062.
25. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, United States (2015). p. 3431–40.
26. Zheng S, Sadeep J, Bernardino R, Vibhav V, Su ZZ, Du DL, et al. Conditional random fields as recurrent neural networks. *Proceedings of the IEEE International Conference on Computer Vision*. Santiago, Chile (2015). p. 1529–37.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, United States. (2016). p. 770–8.
28. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. *Value Health* (2015) 18(2):137–40. doi:10.1016/j.jval.2014.12.005
29. Melendez J, Ginneken BV, Maduskar P, Philipsen RH, Reither K, Breuninger M, et al. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. *IEEE Trans Med Imaging* (2015) 34(1):179–92. doi:10.1109/TMI.2014.2350539
30. Hu P, Wu F, Peng J, Liang P, Kong D. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. *Phys Med Biol* (2016) 61(24):8676. doi:10.1088/1361-6560/61/24/8676
31. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* (2017) 542(7639):115–8. doi:10.1038/nature21056
32. Song Y, Tan EL, Jiang X, Cheng JZ, Ni D, Chen S, et al. Accurate cervical cell segmentation from overlapping clumps in Pap smear images. *IEEE Trans Med Imaging* (2017) 36:288–300. doi:10.1109/TMI.2016.2606380
33. Xing F, Xie Y, Yang L. An automatic learning-based framework for robust nucleus segmentation. *IEEE Trans Med Imaging* (2016) 35(2):550–66. doi:10.1109/TMI.2015.2481436
34. Fu H, Xu Y, Wong DWK, Liu J. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on. IEEE*. Prague, Czech Republic (2016). p. 698–701.
35. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The importance of skip connections in biomedical image segmentation. *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Athens, Greece (2016). p. 179–87.
36. Kamnitsas K, Ledig C, Newcombe VF, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* (2017) 36:61. doi:10.1016/j.media.2016.10.004
37. Tran PV. *A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI*. (2016) arXiv:1604.00494.
38. Ben-Cohen A, Diamant I, Klang E, Amitai M, Greenspan H. Fully convolutional network for liver segmentation and lesions detection. *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Athens, Greece (Vol. 10008) (2016). p. 77–85.
39. Thong W, Kadoury S, Piche N, Pal CJ. Convolutional networks for kidney segmentation in contrast-enhanced CT scans. *Computer Methods in*

Biomechanics and Biomedical Engineering: Imaging & Visualization. Tel Aviv, Israel (2016). p. 1–6.
40. Roth HR, Lu L, Farag A, Sohn A, Summers RM. Spatial aggregation of holistically-nested networks for automated pancreas segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Athens, Greece (2016). p. 451–9.
41. Cheng R, Roth HR, Lu L, Wang S, Turkbey B, Gandler W. Active appearance model and deep learning for more accurate prostate segmentation on MRI. *Med Imaging Image Process* (2016) 9784:97842I. doi:10.1117/12.2216286
42. Cha KH, Hadjiiski LM, Samala RK, Chan HP, Cohan RH, Caoili EM, et al. Bladder cancer segmentation in CT for treatment response assessment: application of deep-learning convolution neural network – a pilot study. *Tomography* (2016) 2:421–9. doi:10.18383/j.tom.2016.00184
43. Xu Y, Li Y, Liu M, Wang Y, Lai M, Chang EIC. Gland instance segmentation by deep multichannel side supervision. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Athens, Greece (2016). p. 496–504.
44. Korez R, Likar B, Pernus F, Vrtovec T. Model-based segmentation of vertebral bodies from MR images with 3D CNNs. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Athens, Greece (2016). p. 433–41.
45. Bibault JE, Giraud P, Burgun A. Big data and machine learning in radiation oncology: state of the art and future prospects. *Cancer Lett* (2016) 382(1): 110–7. doi:10.1016/j.canlet.2016.05.033
46. Cha KH, Hadjiiski L, Samala RK, Chan HP, Caoili EM, Cohan RH. Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. *Med Phys* (2016) 43(4):1882–96. doi:10.1118/1.4944498
47. Hu P, Wu F, Peng J, Bao Y, Feng C, Kong D. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. *Int J Comput Assist Radiol Surg* (2017) 12(3):399–411. doi:10.1007/s11548-016-1501-5
48. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. *Med Phys* (2017) 44(4):1408–19. doi:10.1002/mp.12155
49. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* (2017) 44(2):547–57. doi:10.1002/mp.12045
50. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: convolutional architecture for fast feature embedding. *ArXiv* (1408) 2014:5093.
51. Chetlur S, Woolley C, Vandermersch P, Cohen J, Tran J, Catanzaro B, et al. cuDNN: efficient primitives for deep learning. *ArXiv* (1410) 2014:0759.
52. Leunens G, Menten J, Weltens C, Verstraete J, Schueren EVD. Quality assessment of medical decision making in radiation oncology: variability in target volume delineation for brain tumours. *Radiother Oncol* (1993) 29(2):169–75. doi:10.1016/0167-8140(93)90243-2
53. Lu L, Cuttino L, Barani I, Song S, Fatyga M, Murphy M, et al. SU-FF-J-85: inter-observer variation in the planning of head/neck radiotherapy. *Med Phys* (2006) 33(6):2040. doi:10.1118/1.2240862
54. Caravatta L, Macchia G, Mattiucci GC, Sainato A, Cernusco NL, Mantello G, et al. Inter-observer variability of clinical target volume delineation in radiotherapy treatment of pancreatic cancer: a multi-institutional contouring experience. *Radiat Oncol* (2014) 9(1):198. doi:10.1186/1748-717X-9-198
55. Chen A, Deeley MA, Niermann KJ, Moretti L, Dawant BM. Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images. *Med Phys* (2010) 37(12):6338–46. doi:10.1118/1.3515459
56. Jean-François D, Andreas B. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. *Radiat Oncol* (2013) 8(1):1–11. doi:10.1186/1748-717X-8-154
57. Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: a feature-driven model-based approach. *Med Phys* (2011) 38(11):6160–70. doi:10.1118/1.3654160
58. Tsuji SY, Hwang A, Weinberg V, Yom SS, Quivey JM, Xia P. Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. *Int J Radiat Oncol Biol Phys* (2010) 77(3):707–14. doi:10.1016/j.ijrobp.2009.06.012
59. Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, Mcdonald MW, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *Int J Radiat Oncol Biol Phys* (2010) 77(3):959–66. doi:10.1016/j.ijrobp.2009.09.023
60. Gorthi S, Duay V, Houhou N, Cuadra MB, Schick U, Becker M. Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. *IEEE J Sel Topics Signal Process* (2009) 3(1):135–47. doi:10.1109/JSTSP.2008.2011104
61. Yang J, Beadle BM, Garden AS, Schwartz DL, Aristophanous M. A multimodality segmentation framework for automatic target delineation in head and neck radiotherapy. *Med Phys* (2015) 42(9):5310–20. doi:10.1118/1.4928485
62. Yang J, Beadle BM, Garden AS, Gunn B, Rosenthal D, Ang K, et al. Auto-segmentation of low-risk clinical target volume for head and neck radiation therapy. *Pract Radiat Oncol* (2014) 4(1):e31–7. doi:10.1016/j.prro.2013.03.003
63. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging* (1994) 13(4):716–24. doi:10.1109/42.363096

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Architecture of Deep Deconvolutional Neural Network (DDNN)

As is shown in **Figure 2**, the architecture of the proposed DDNN consisted of two parts, each of which had its own role. The encoder networks consisted of 13 convolutional layers for feature extraction. All the kernels of convolutional layers had a window size of 3 × 3, a stride of 1, and a padding of 1 pixel. In addition, there was a batch normalized option following each convolution layer and then an element-wise rectified linear non-linearity max (0, x) was applied. The pooling options were added after the layers of conv1_2, conv2_2, conv3_3, conv4_3, conv5_3, de_conv5_1, and de_conv4_1 in order to get the robust feature. Specifically, the input size of the medical images in this work was cropped to 417 × 417 with 3 channels. Conv1_1 and conv1_2 convolved the input to 417 × 417 × 64, and then reduced to 209 × 209 × 64 feature maps using pooling option with kernel size of 3 × 3, a stride of 2, and a padding of 1 pixel. Similarly, the layers of conv2_1 and conv2_2 took pool1 as input. After using 3 × 3 convolution with a stride of 1 and a padding of 1 pixel, it produced 105 × 105 × 256 feature maps and then was pooled by pool2 and convolved by conv3, conv4, and conv5. The max pooling options of pool4 and pool5 with 3 × 3 filter size, pad 1 and stride 1, resulted in a 53 × 53 × 512 output. The pooling options reduced the spatial size of feature map, so the feature map needed to be recovered to the original spatial size for segmentation task. Most previous methods used bilinear interpolation to get high-resolution image; however, a coarse segmentation was not enough to produce good performance for nasopharyngeal cancer. Therefore, the decoder part deployed a deep deconvolution neural network which took pool5 as input and a serial of deconvolution layers for upsampling. All the deconvolutional layers used 3 × 3 convolution with the padding size of 1. At de_conv5_3, de_conv4_3, and de_conv3_3, the stride was set to 2. For others, the stride was set to 1. After 8× enlarging, the feature maps recovered the high resolution as same as input. At fc6 and fc7 layers, we replaced fully connected layer with 1 × 1 convolution. Thus, we can carry on the pixel-level classification for the segmentation task. The final outputs generated predicted label for each pixel.

frontiers in Oncology

ORIGINAL RESEARCH
published: 05 March 2018
doi: 10.3389/fonc.2018.00035

Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia

Hubert S. Gabryś[1,2,3], Florian Buettner[4], Florian Sterzing[3,5,6], Henrik Hauswald[3,5,6] and Mark Bangert[1,3]**

[1]*Department of Medical Physics in Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany,* [2]*Medical Faculty of Heidelberg, Heidelberg University, Heidelberg, Germany,* [3]*Heidelberg Institute for Radiation Oncology (HIRO), Heidelberg, Germany,* [4]*Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany,* [5]*Clinical Cooperation Unit Radiation Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany,* [6]*Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany*

OPEN ACCESS

Edited by:
Issam El Naqa,
University of Michigan, United States

Reviewed by:
John C. Roeske,
Loyola University Medical Center, United States
John Austin Vargo,
West Virginia University Hospitals, United States

***Correspondence:**
Hubert S. Gabryś
h.gabrys@dkfz.de;
Mark Bangert
m.bangert@dkfz.de

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *21 November 2017*
Accepted: *01 February 2018*
Published: *05 March 2018*

Citation:
Gabryś HS, Buettner F, Sterzing F, Hauswald H and Bangert M (2018) Design and Selection of Machine Learning Methods Using Radiomics and Dosiomics for Normal Tissue Complication Probability Modeling of Xerostomia. Front. Oncol. 8:35. doi: 10.3389/fonc.2018.00035

Purpose: The purpose of this study is to investigate whether machine learning with dosiomic, radiomic, and demographic features allows for xerostomia risk assessment more precise than normal tissue complication probability (NTCP) models based on the mean radiation dose to parotid glands.

Material and methods: A cohort of 153 head-and-neck cancer patients was used to model xerostomia at 0–6 months (early), 6–15 months (late), 15–24 months (long-term), and at any time (a longitudinal model) after radiotherapy. Predictive power of the features was evaluated by the area under the receiver operating characteristic curve (AUC) of univariate logistic regression models. The multivariate NTCP models were tuned and tested with single and nested cross-validation, respectively. We compared predictive performance of seven classification algorithms, six feature selection methods, and ten data cleaning/class balancing techniques using the Friedman test and the Nemenyi *post hoc* analysis.

Results: NTCP models based on the parotid mean dose failed to predict xerostomia (AUCs < 0.60). The most informative predictors were found for late and long-term xerostomia. Late xerostomia correlated with the contralateral dose gradient in the anterior–posterior (AUC $= 0.72$) and the right–left (AUC $= 0.68$) direction, whereas long-term xerostomia was associated with parotid volumes (AUCs > 0.85), dose gradients in the right–left (AUCs > 0.78), and the anterior–posterior (AUCs > 0.72) direction. Multivariate models of long-term xerostomia were typically based on the parotid volume, the parotid eccentricity, and the dose–volume histogram (DVH) spread with the generalization AUCs ranging from 0.74 to 0.88. On average, support vector machines and extra-trees were the top performing classifiers, whereas the algorithms based on logistic regression were the best choice for feature selection. We found no advantage in using data cleaning or class balancing methods.

Conclusion: We demonstrated that incorporation of organ- and dose-shape descriptors is beneficial for xerostomia prediction in highly conformal radiotherapy treatments. Due to strong reliance on patient-specific, dose-independent factors, our results underscore the need for development of personalized data-driven risk profiles for NTCP models of xerostomia. The facilitated machine learning pipeline is described in detail and can serve as a valuable reference for future work in radiomic and dosiomic NTCP modeling.

Keywords: radiotherapy, IMRT, NTCP, xerostomia, head and neck, machine learning, radiomics, dosiomics

1. INTRODUCTION

Radiotherapy is the main treatment for head-and-neck tumors. Incidental irradiation of salivary glands often impairs their function, causing dryness in the mouth (xerostomia). Xerostomia significantly reduces patients' quality of life, leading to dental health deterioration, oral infections, and difficulties in speaking, chewing, and swallowing.

The Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) group recommended sparing at least one parotid gland to a mean dose $<$20 Gy or both parotid glands to a mean dose $<$25 Gy (1). Large-cohort studies confirmed that the mean dose is a good predictor of xerostomia (2, 3). However, it has also been observed that the mean dose failed to recognize patients at risk in cohorts where the majority of patients had met the QUANTEC guidelines, although the prevalence of xerostomia was reduced (4–6).

In recent years, a number of studies have investigated various patient- and therapy-related factors in hope of more precise xerostomia predictions. These included the mean dose to submandibular glands and the oral cavity (5, 7–9), sparing of the parotid stem cells region (10), three-dimensional dose moments (4), CT image features (11, 12), patients' T stage, age, financial status, education, smoking, etc. (4, 5, 8).

Moreover, there has been growing interest in the adoption of machine learning classifiers in NTCP modeling (13–15). Buettner et al. used Bayesian logistic regression together with dose-shape features to predict xerostomia in head-and-neck cancer patients (4). Support vector machines were employed to model radiation-induced pneumonitis (16). Ospina et al. predicted rectal toxicity following prostate cancer radiotherapy using random forests (17).

Nevertheless, despite the growing interest in data-driven methods, there have been no published studies so far systematically evaluating how different machine learning techniques can be used to address the challenges specific to NTCP modeling. These include class imbalance due to low prevalence rates, heterogeneous and noisy data, large feature spaces, irregular follow-up times, etc. A comparable work has already been presented in the fields of bioinformatics (18, 19) and radiomics (20). Such analysis is missing for NTCP modeling, although it seems especially relevant.

In this context, we examined associations between xerostomia and various features describing parotid shape (radiomics), dose shape (dosiomics), and demographic characteristics. Besides investigating the individual predictive power of the features, we comprehensively evaluated the suitability of seven machine learning classifiers, six feature selection methods, and ten data cleaning/class balancing algorithms for multivariate NTCP modeling. The obtained results were compared to mean-dose models and the morphological model proposed by Buettner et al. (4). Furthermore, we proposed a longitudinal approach for NTCP modeling that includes the time after treatment as a model covariate. Doing so, rather than binning the data around a certain time point, better reflects the underlying data due to often irregular follow-up times.

2. MATERIALS AND METHODS

2.1. Patients

The retrospective patient cohort collected for this study comprised head-and-neck cancer patients treated with radiotherapy at Heidelberg University Hospital in years 2010–2015. After excluding patients with nonzero baseline xerostomia, replanning during the treatment, tumor in the parotid gland, second irradiation, second chemotherapy, or ion beam boost, the cohort consisted of 153 patients. Patient and tumor characteristics are listed in **Table 1**. The study was approved by the Ethics Committee of Heidelberg University.

2.2. End Points

For this study, we analyzed 693 xerostomia toxicity follow-up reports. We aimed to model moderate-to-severe xerostomia defined as grade 2 or higher according to Common Terminology Criteria for Adverse Effects (CTCAE) v4.03 (21). In 74% of cases, either CTCAE v3.0 or v4.03 grading scale was used. Dry mouth (xerostomia) definitions were the same in both versions so no inconsistency in grading was introduced. In case no score was provided but descriptive toxicity information was available, appropriate scores were assigned together with Heidelberg University Hospital clinicians. To minimize intra- and interobserver variability in this process, a set of rules in the form of a dictionary was introduced.

The follow-up reports were collected, on average, at 3-month intervals (**Figure 1**). The number of toxicity evaluations and the length of the follow-up varied from patient to patient. Due to the time-characteristic and the irregularity of the follow-up, two approaches were taken to model xerostomia: a time-specific approach and a longitudinal approach. In the time-specific approach, three time intervals were defined: 0–6, 6–15, and 15–24 months, to investigate early, late, and long-term xerostomia, respectively. In case there were multiple follow-up

TABLE 1 | Patients and tumor characteristics.

	All	0–6 months			6–15 months			15–24 months		
		Grade 0	Grade 1	Grade 2	Grade 0	Grade 1	Grade 2	Grade 0	Grade 1	Grade 2
Total patients	153	17	87	30	19	99	13	15	53	9
Age										
Median	61	60	60	62	60	61	61	61	61	61
Q1–Q3	55–66	54–66	54–64	53–69	57–63	53–66	54–68	55–68	52–66	54–68
Range	29–82	44–78	29–82	43–80	49–75	29–82	43–74	47–80	39–78	41–80
Sex										
Female	37	5	19	7	6	24	2	2	9	4
Male	116	12	68	23	13	75	11	13	44	5
Tumor site										
Hypopharynx/larynx	37	7	20	7	7	20	2	3	15	0
Nasopharynx	12	0	8	2	2	8	1	0	5	0
Oropharynx	99	9	57	20	10	69	9	11	32	9
Other	5	1	2	1	0	2	1	1	1	0
Radiation modality										
IMRT	37	2	25	5	1	29	2	2	18	1
Tomotherapy	116	15	62	25	18	70	11	13	35	8
Ipsi parotid dose (Gy)										
Median	24.3	22.9	25.0	23.0	19.5	24.8	25.9	22.9	23.8	24.5
Q1–Q3	20.6–27.6	18.5–24.6	21.4–29.0	21.4–25.4	16.8–24.3	21.8–28.7	21.8–27.2	18.5–31.5	20.8–26.4	21.6–26.2
Range	0.4–63.4	0.4–36.0	7.4–61.4	4.6–59.0	0.4–32.9	4.6–61.4	17.3–63.4	0.4–51.4	4.6–46.0	17.3–63.4
Contra parotid dose (Gy)										
Median	19.9	19.4	20.3	19.6	15.6	20.5	20.4	12.7	19.7	20.1
Q1–Q3	15.4–23.1	13.1–21.8	15.2–23.8	16.5–22.0	10.3–20.7	16.3–23.8	19.8–23.1	5.2–17.9	16.3–23.7	16.4–22.3
Range	0.3–30.9	0.3–24.9	4.1–28.6	4.2–26.2	0.3–27.9	4.1–30.9	15.1–26.2	0.3–27.9	4.1–27.2	15.1–26.0

The total number of patients differs among the groups due to the follow-up availability.

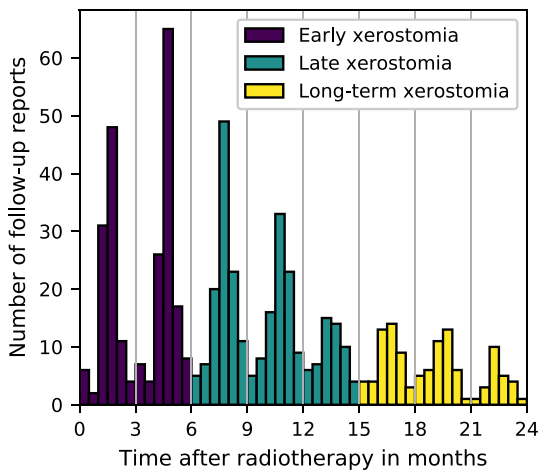

FIGURE 1 | Frequency of the follow-up reports collection.

reports available for individual patients, the final toxicity score was calculated as the arithmetic mean rounded to the nearest integer number with x.5 being rounded up. In the longitudinal approach, no time-intervals were defined and no toxicity grades were averaged. Instead, each patient evaluation served as a separate observation and the time after treatment was included as a covariate in the model.

2.3. Features

The candidate xerostomia predictors comprised demographic, radiomic, and dosiomic features (**Table 2**). The radiomic and the dosiomic features were extracted from the CT- and the dose-cubes read from treatment planning DICOM files. In a preprocessing step, all the cubes were linearly interpolated to an isotropic 1 mm resolution. Moreover, we wanted to analyze the features in terms of ipsi- and contralateral rather than left and right parotid glands. This would, however, mean that certain spatial features would

TABLE 2 | Feature sets before and after the removal of highly correlated pairs (Kendall's $|\tau| > 0.5$).

Feature group	Initial feature set	Final feature set
Demographics	Age, sex	Age, sex
Parotid shape	Volume, area, sphericity, eccentricity, compactness, λ_1, λ_2, λ_3	Volume, sphericity, eccentricity
Dose–volume histogram	Mean, spread, skewness, D2, D98, D10, D20, D30, D40, D50, D60, D70, D80, D90, V10, V15, V20, V25, V30, V35, V40, V45, entropy, uniformity	Mean, spread, skewness
Subvolume mean dose	$s_x^1, s_x^2, s_x^3, s_y^1, s_y^2, s_y^3, s_z^1, s_z^2, s_z^3$	
Spatial dose gradient	Gradient_x, gradient_y, gradient_z	Gradient_x, gradient_y, gradient_z
Spatial dose spread	$\eta_{200}, \eta_{020}, \eta_{002}$	$\eta_{200}, \eta_{020}, \eta_{002}$
Spatial dose correlation	$\eta_{110}, \eta_{101}, \eta_{011}$	$\eta_{110}, \eta_{101}, \eta_{011}$
Spatial dose skewness	$\eta_{300}, \eta_{030}, \eta_{003}$	$\eta_{300}, \eta_{030}, \eta_{003}$
Spatial dose coskewness	$\eta_{012}, \eta_{021}, \eta_{120}, \eta_{102}, \eta_{210}, \eta_{201}$	$\eta_{012}, \eta_{021}, \eta_{120}, \eta_{102}, \eta_{210}, \eta_{201}$

Feature definitions are provided in Appendix A.

have either positive or negative value, depending on the tumor location (left or right). In order to solve that issue, the cubes were flipped through the sagittal plane for cases with the mean dose to the right parotid gland higher than the mean dose to the left parotid gland. All feature definitions were based on the LPS coordinate system, that is (right to left, anterior to posterior, inferior to superior). The detailed definitions of the features are provided in Appendix A.

To reduce feature redundancy, the Kendall rank correlation coefficient was calculated for all feature pairs. Kendall's τ allows to measure ordinal association between two features, that is agreement in ranks assigned to the observations. It can be interpreted as a difference between the probability that both features rank a random pair of observations in the same way and the probability that they rank these observations in a different way (22). We considered feature pairs with $|\tau| > 0.5$ in both glands as highly correlated and suitable for rejection from the feature set. This arbitrarily chosen threshold corresponds to a 75% probability that the two features rank a random pair of observations in the same way. Whenever a pair of features was found highly correlated, we decided to keep the feature that was conceptually and computationally simpler, e.g., mean dose over Dx, parotid volume over parotid compactness, etc.

2.4. Previously Proposed NTCP Models

Logit and probit NTCP models based on the mean dose to parotid glands have been extensively used in modeling xerostomia (2, 3, 23, 24). We have tested four different mean-dose models to evaluate predictive power of the mean dose in our cohort: three univariate logistic regression models based on the ipsilateral mean dose, the contralateral mean dose, and the mean dose to both parotid glands, as well as one bivariate logistic regression model based on the mean dose to contralateral and to ipsilateral parotid glands.

As an alternative to the mean-dose models, Buettner et al. (4) proposed a multivariate logistic regression model based on three-dimensional dose moments to predict xerostomia. The model was retrained and tested on our data set.

2.5. Univariate Analysis

The univariate analysis was performed to investigate associations of single features with the outcome at different time intervals. First, all features were normalized *via* Z-score normalization to zero mean and unit variance. Next, for each feature, the Mann–Whitney U statistic was calculated. The area under the receiver operating characteristic curve (AUC) is directly related to the U statistic and follows from the formula $AUC = \frac{U}{n_- n_+}$, where n_- and n_+ are the size of the negative and the size of the positive class, respectively (25). For all AUCs, 95% confidence intervals were estimated by bias-corrected and accelerated (BCa) bootstrap (26). The number of type I errors, that is falsely rejected null hypotheses, was controlled with the false discovery rate (FDR). The FDR is defined as the expected proportion of true null hypotheses in the set of all the rejected hypotheses (27). We applied the Gavrilov-Benjamini-Sarkar procedure to bound the FDR ≤ 0.05 (28). Additionally, for each feature, univariate logistic regression models were fitted and tolerance values corresponding to 20% (TV20), 10% (TV10), and 5% (TV5) complication probability were calculated.

2.6. Multivariate Analysis

The multivariate analysis allowed to examine interactions between the features and their relative relevancy and redundancy. It was a multi-step process comprising feature-group selection, feature scaling, sampling (data cleaning and/or class balancing), feature selection, and classification. The workflow is presented in **Figure 2**.

2.6.1. Workflow

The first step of the workflow was a random selection of the feature-groups (**Table 2**) used for model training. It allowed for an initial, unsupervised dimensionality reduction of the feature space, which typically translates into an improved predictive performance and a more straightforward interpretation of the models. The selection was realized by performing a Bernoulli trial for every feature group with a 50% chance of success. If

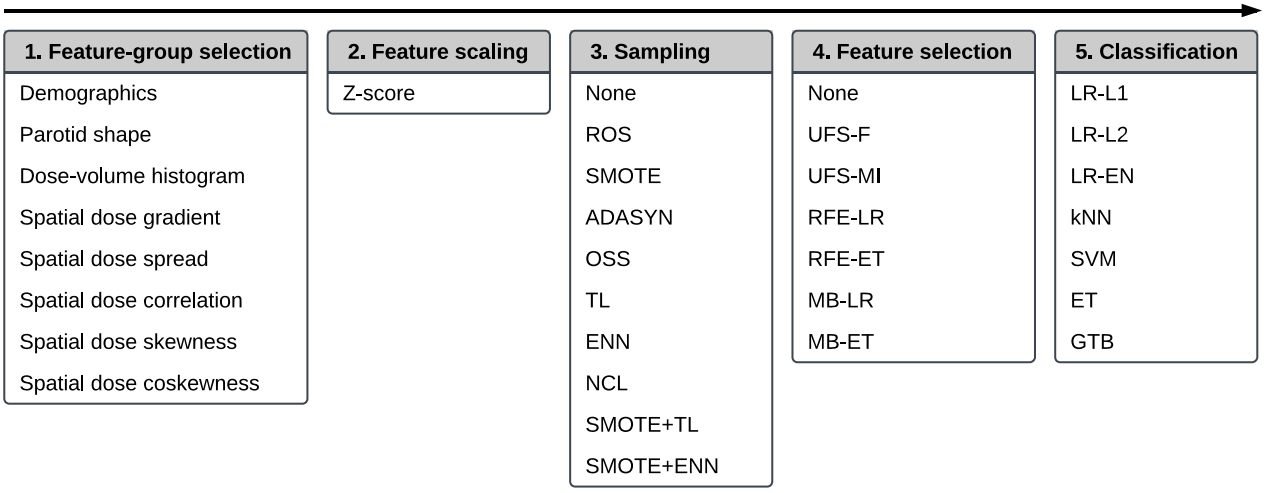

FIGURE 2 | The workflow of a multivariate five-step model building comprising, in this order, feature-group selection, feature scaling, sampling, feature selection, and classification.

TABLE 3 | Predictive performance of the mean-dose models and the morphological model proposed by Buettner et al. (4), that is logistic regression with η^{i}_{111}, η^{c}_{002}, η^{c}_{300}, and $\eta^{i}_{110}\eta^{c}_{110}$.

End point	Model	AUC
Early	Mean^{i}	0.58 (0.56–0.60)
	Mean^{c}	0.42 (0.41–0.44)
	Mean^{b}	0.50 (0.48–0.53)
	Mean^{i}, mean^{c}	0.49 (0.48–0.51)
	Morphological	0.42 (0.40–0.44)
Late	Mean^{i}	0.48 (0.44–0.51)
	Mean^{c}	0.58 (0.55–0.61)
	Mean^{b}	0.55 (0.52–0.58)
	Mean^{i}, mean^{c}	0.54 (0.51–0.57)
	Morphological	0.59 (0.56–0.62)
Long-term	Mean^{i}	0.40 (0.37–0.44)
	Mean^{c}	0.58 (0.55–0.61)
	Mean^{b}	0.56 (0.52–0.60)
	Mean^{i}, mean^{c}	0.47 (0.44–0.50)
	Morphological	0.64 (0.60–0.67)
Longitudinal	Mean^{i}	0.51 (0.45–0.56)
	Mean^{c}	0.57 (0.51–0.62)
	Mean^{b}	0.50 (0.44–0.55)
	Mean^{i}, mean^{c}	0.52 (0.46–0.58)
	Morphological	0.55 (0.49–0.60)

i, ipsilateral gland; c, contralateral gland; b, both glands.

a given group was selected, all features belonging to this group were accepted for further analysis. If no group was selected after performing all Bernoulli trials, the procedure was repeated for all feature groups.

In the second step, all features were scaled *via* Z-score normalization. Normalization of the features often improves stability and speed of optimization algorithms.

The third step served the purpose of class balancing and data cleaning. A class imbalance, noise, and a small size of the minority class can negatively affect the performance of a predictive model (29, 30). We investigated whether sampling methods designed to reduce noise and improve definitions of class clusters could enhance model performance. Ten algorithms were examined: random oversampling (ROS), synthetic minority oversampling (SMOTE), adaptive synthetic sampling (ADASYN), one-sided selection (OSS), Tomek links (TL), the Wilson's edited nearest neighbor rule (ENN), the neighborhood cleaning rule (NCL), synthetic minority oversampling followed by the Wilson's edited nearest neighbor rule (SMOTE + ENN), and synthetic minority oversampling followed by Tomek links (SMOTE + TL). The detailed description of the sampling algorithms is given in Appendix B.

The fourth step of the analysis was feature selection. The rationale for feature selection is a reduction of model complexity, which facilitates understanding of the relations between the predictors and the modeled outcome (here: xerostomia) (31). In this study, we tested six feature selection algorithms: univariate feature selection by F-score (UFS-F), univariate feature selection by mutual information (UFS-MI), recursive feature elimination by logistic regression (RFE-LR), recursive feature elimination by extra-trees (RFE-ET), model-based feature selection by logistic regression (MB-LR), and model-based feature selection by extra-trees (MB-ET). The details on the feature selection algorithms are provided in Appendix C.

The last step of the workflow was classification. We compared seven classification algorithms: logistic regression with L1 penalty (LR-L1), logistic regression with L2 penalty (LR-L2), logistic regression with elastic net penalty (LR-EN), k-nearest neighbors (kNN), support vector machines (SVM), extra-trees (ET), and gradient tree boosting (GTB). A more detailed description of the classification algorithms is given in Appendix D.

The models were build for every combination of the classification, feature selection, and sampling algorithms. This resulted in 490 models per end point or 1,960 models in total. A given classifier or a feature selection algorithm was involved in 210 time-specific and 70 longitudinal models. Every sampling method was part of 147 time-specific and 49 longitudinal models.

2.6.2. Model Tuning

In the process of model building every model was tuned, that is its hyperparameters were optimized to maximize the prediction performance. The type and the range of the hyperparameters were based on previously reported values that worked well in various machine learning tasks (Appendices B, C, and D).

For each model, the hyperparameter optimization was realized by a random search (32). First, 300 random samples were selected from the hyperparameter space. Secondly, for each hyperparameter sample, the model performance was evaluated using cross-validation. Lastly, the model was retrained using all data with the hyperparameter configuration that maximized the cross-validated AUC.

In the time-specific models, the cross-validation was done by the stratified Monte Carlo cross-validation (MCCV) (33) with 300 splits and 10% of observations held out for testing at each split. For the longitudinal models, we used modified leave-pair-out cross-validation (LPOCV) (34, 35). In our LPOCV implementation, all the training observations sharing patient ID with the test fold observations were removed at each split. This decision was motivated by the fact that the observations sharing patient ID differ only in the time of the follow-up evaluation; not removing them from the training fold would lead to overoptimistic performance scores. Additionally, instead of all possible positive–negative pairs, as in typical LPOCV, only a random subset of 300 positive–negative pairs was used. This allowed for a reduction of the computation time. Confidence intervals for the model tuning AUC estimates were calculated with BCa bootstrap.

2.6.3. Comparison of Machine Learning Algorithms

In order to compare the algorithms in terms of their influence on the average predictive performance of the model, we looked at the classifiers, the feature selection algorithms, and the sampling methods separately. Additionally, the analysis was performed independently for the time-specific and the longitudinal models.

The statistical significance of the differences between the algorithms was evaluated by the Friedman test followed by the Nemenyi *post hoc* analysis. The Friedman test computes average performance ranks of the algorithms and tests whether they have the same influence on the AUC score of the model. If the null

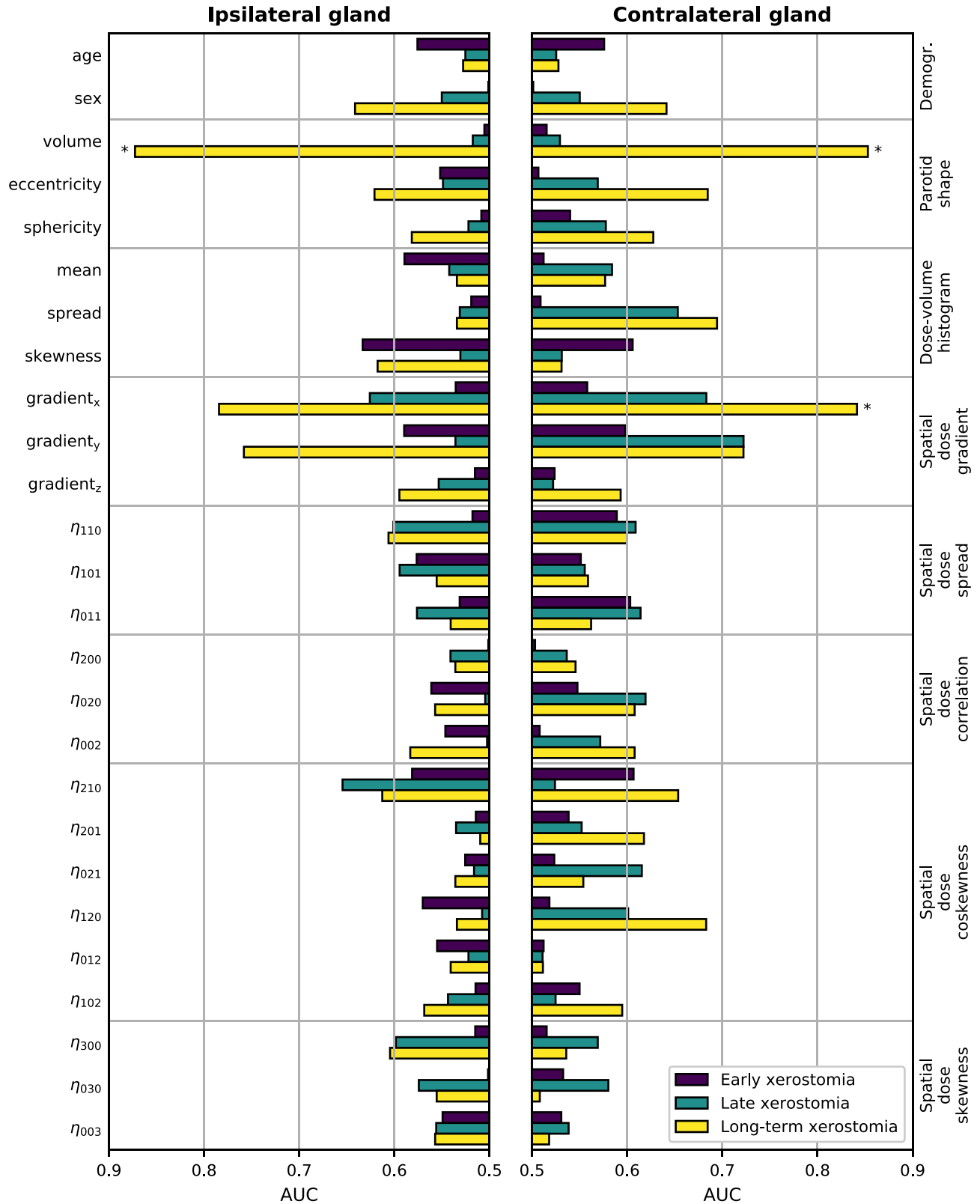

FIGURE 3 | Predictive power of individual features in the time-specific models measured with the area under the receiver operating characteristic curve (AUC). The left-hand side vertical axis lists the features, the right-hand side vertical axis lists the feature groups. The AUCs were calculated from the corresponding Mann–Whitney *U* statistic. Bars marked with * are significant at the false discovery rate (FDR) ≤ 0.05.

hypothesis was rejected, we proceeded with the *post hoc* analysis. With the Nemenyi *post hoc* test, we calculated the critical difference at a significance level of 0.05. When the average performance ranks of two algorithms differed by at least the critical difference, they were significantly different.

As mentioned before, this analysis was repeated six times to test the classifiers, the feature selection algorithms, and the sampling methods separately in the time-specific and the longitudinal models. Therefore, the Holm–Bonferroni method was used to control the family-wise error rate (FWER) of the Friedman tests, that is the probability of making at least one incorrect rejection of a true null hypothesis in any of the comparisons (36). The significance level for the FWER was set to 0.05.

2.6.4. Generalization Performance

Hyperparameter optimization comes at a cost. On the one hand, it allows to tune the model so it fits well the underlying data. On the other hand, the performance of the tuned model may be overoptimistic due to a favorable selection of hyperparameters. In order to estimate the generalization performance of a model, that is its performance on new, unseen data, the data used for model tuning must be separate from the data used for model testing. Due to the modest size of our data set, instead of dividing the data to training, validation, and test folds, we decided to test the models using nested-cross validation (37).

Nested cross-validation is essentially cross validation within cross validation. Part of the data is set aside for testing and the rest is used for model tuning (as described in the previous section). Next, the tuned model is tested on the part of data previously set aside for testing. Then, the procedure is repeated, that is another randomly selected part of the data is set aside for testing and the rest is used for model tuning. This is repeated until the desired number of iterations is achieved.

Unfortunately, due to high computation cost, it was not feasible to calculate the expected generalization performance of all 1,960 models. Therefore, the models were first stratified by end point and classifier, and then nested cross-validation was conducted for the best performing models. The inner loops of the nested cross-validation, which were responsible for model tuning, were the same as described in Section 2.6.2. The outer loops were realized by the MCCV with 100 splits and a 10% test fold (time-specific models) or the modified LPOCV (longitudinal models). Confidence intervals for the generalization AUCs were calculated with BCa bootstrap.

2.7. Software

The MATLAB code used for DICOM import, processing, and feature extraction was made publicly available on GitHub (https://github.com/hubertgabrys/DicomToolboxMatlab). For visualization, statistical analysis, model building, and model testing, the following open-source Python packages were used: imbalanced-learn (38), Matplotlib (39), NumPy & SciPy (40), Orange (41), Pandas (42), scikit-learn (43), scikits-bootstrap, and XGBoost (44).

3. RESULTS

3.1. Feature Correlations

After removing the features correlated with the mean dose, the skewness of the dose–volume histogram, and the parotid volume, there were no highly correlated feature pairs left. The remaining features are listed in **Table 2**.

3.2. Mean-Dose and Morphological Models

The predictive performance scores of the mean-dose models and the morphological model are presented in **Table 3**. The mean-dose models failed to predict xerostomia (AUC < 0.60) at all time-intervals as well as in the longitudinal approach. The morphological model achieved fair performance (AUC $= 0.64$) only in predicting long-term xerostomia.

3.3. Univariate Analysis

The results of the univariate analysis are presented in **Figure 3**. There was little association between single predictors and xerostomia within the first six months after treatment. Late xerostomia

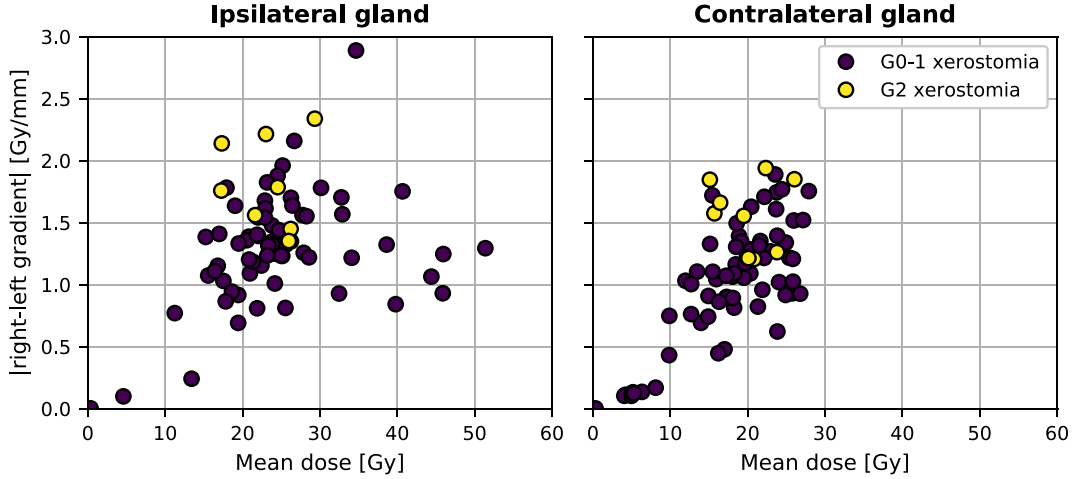

FIGURE 4 | The mean dose and the absolute right–left dose gradient distribution in our patient cohort.

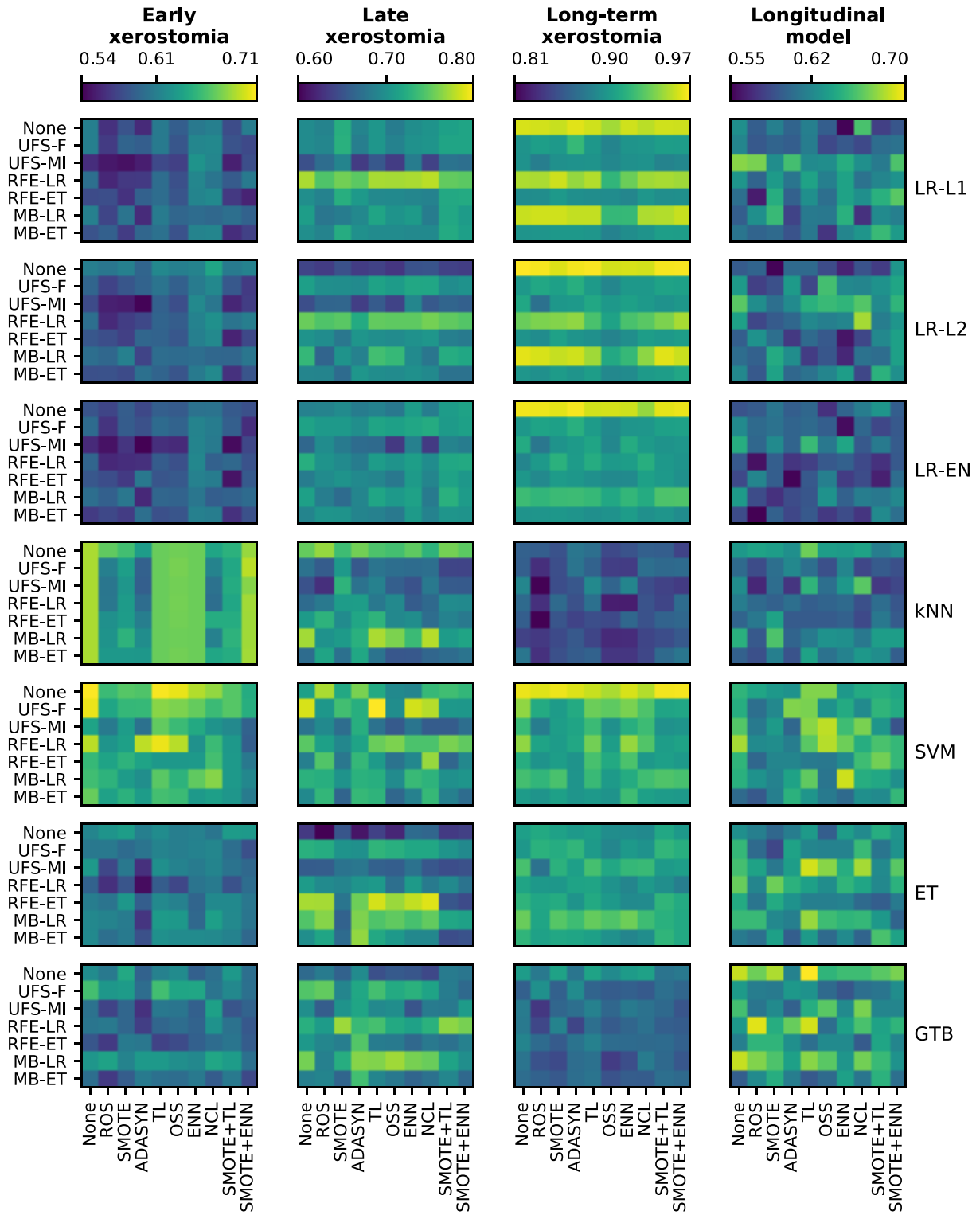

FIGURE 5 | A comparison of classification, feature selection, and sampling algorithms in terms of their predictive performance in model tuning. All heat maps in a given column belong to a single end point, whereas all heat maps in a given row correspond to a single classifier. In each heat map, rows represent feature selection algorithms and columns correspond to sampling methods. The color maps are normalized per end point. The color bar ticks correspond to the worst, average, and the best model performance.

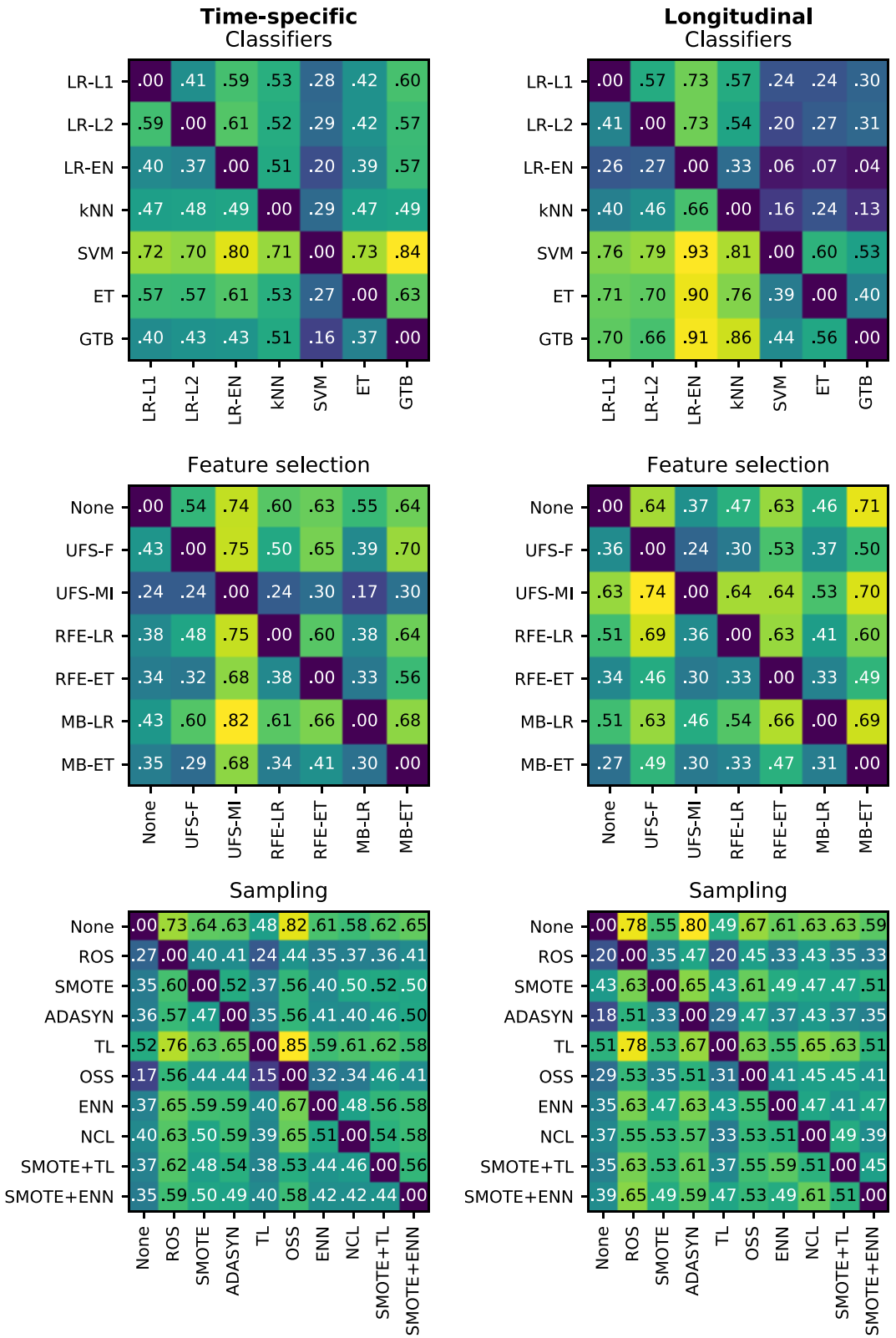

FIGURE 6 | Heat maps showing a proportion of times a given algorithm on the vertical axis outperformed another algorithm on the horizontal axis in terms of the best AUC in model tuning. For example, support vector machines (SVM) performed better than extra-trees (ET) in 73% of the time-specific models.

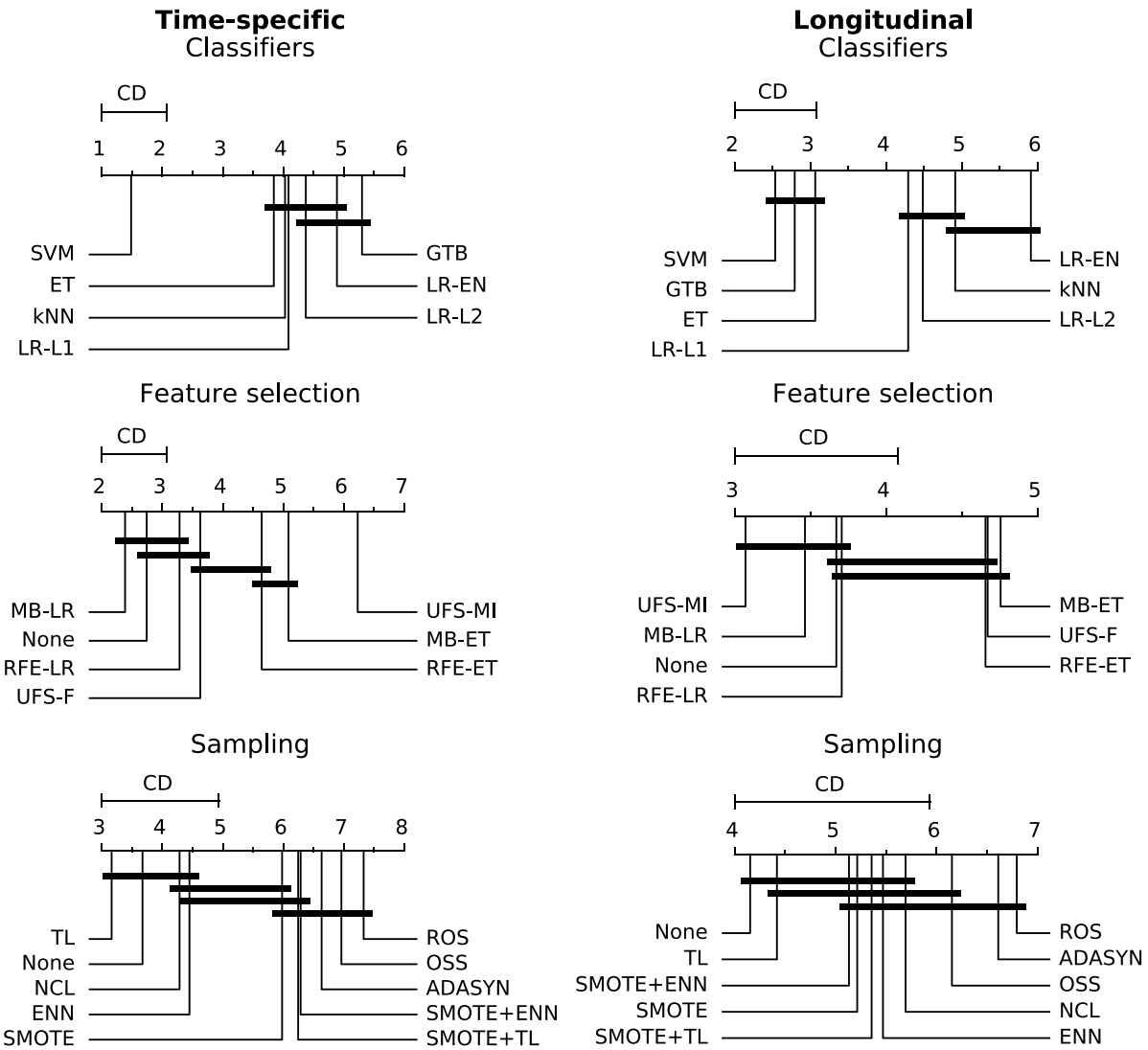

FIGURE 7 | A comparison of classification, feature selection, and sampling methods against one another with the Nemenyi test. Lower ranks correspond to better performance of the algorithm, that is rank 1 is the best. Algorithms which ranks differ by less than the critical difference (CD) are not significantly different at 0.05 significance level and are connected by the black bars.

correlated with individual features slightly better. The most informative were contralateral dose gradients in the right–left direction (AUC = 0.68 (0.53–0.82)) and the anterior–posterior direction (AUC = 0.72 (0.58–0.84)). Nevertheless, the AUCs were too low to be statistically significant at the FDR $\leq$ 0.05. Long-term xerostomia was predicted well by parotid volumes, right–left dose gradients, and anterior–posterior dose gradients. Three models were statistically significant at the FDR $\leq$ 0.05: the ipsilateral parotid volume (AUC = 0.87 (0.75–0.95), TV20 = 9,894 mm^3, TV10 = 15,681 mm^3, TV5 = 21,014 mm^3), the contralateral parotid volume (AUC = 0.85 (0.66–0.98), TV20 = 9,169 mm^3, TV10 = 14,533 mm^3, TV5 = 19,475 mm^3), and the contralateral gradient in the right–left direction (AUC = 0.84 (0.71–0.93), TV20 = 1.49 Gy/mm, TV10 = 1.29 Gy/mm, TV5 = 1.10 Gy/mm). Statistical significance of three tests at the FDR $\leq$ 0.05 translates into a 85.7% and a 99.3% lower bound on the probability that all three tests are truly positive or that at most one test is falsely positive, respectively.

Neither the mean dose to the contralateral nor the mean dose to the ipsilateral parotid gland discriminated well between patients with and without xerostomia in the time-specific and the longitudinal approach. **Figure 4** shows the comparison between the mean dose and the absolute right–left dose gradient values for the patients with long-term xerostomia.

3.4. Comparison of Classification, Feature Selection, and Sampling Algorithms

There was a clear difference in the average performance between early (AUC $\approx$ 0.60), late (AUC $\approx$ 0.70), and long-term (AUC $\approx$ 0.90) xerostomia models (**Figure 5**). After applying the

TABLE 4 | Expected generalization performance of selected models evaluated by nested cross-validation.

End point	Classifier	Feature selection	Sampling	AUC tuning	AUC testing
Early	LR-L1	RFE-ET	NCL	0.62 (0.60–0.64)	0.56 (0.53–0.60)
	LR-L2	RFE-LR	NCL	0.62 (0.60–0.64)	0.46 (0.42–0.49)
	LR-EN	MB-ET	NCL	0.62 (0.60–0.64)	0.54 (0.50–0.57)
	kNN	UFS-F	SMOTE + ENN	0.68 (0.66–0.70)	0.65 (0.62–0.68)[a]
	SVM	UFS-F	None	0.70 (0.68–0.72)	0.57 (0.53–0.61)
	ET	MB-LR	NCL	0.63 (0.61–0.65)	0.44 (0.41–0.47)
	GTB	UFS-F	None	0.66 (0.64–0.68)	0.55 (0.51–0.59)
Late	LR-L1	RFE-LR	NCL	0.78 (0.75–0.80)	0.63 (0.56–0.69)
	LR-L2	RFE-LR	NCL	0.76 (0.73–0.78)	0.60 (0.53–0.66)
	LR-EN	MB-LR	SMOTE + TL	0.73 (0.70–0.76)	0.56 (0.51–0.62)
	kNN	MB-LR	NCL	0.78 (0.76–0.80)	0.62 (0.57–0.67)
	SVM	UFS-F	TL	0.80 (0.77–0.82)	0.52 (0.46–0.58)
	ET	RFE-ET	NCL	0.78 (0.75–0.80)	0.55 (0.50–0.61)
	GTB	MB-LR	OSS	0.77 (0.75–0.79)	0.65 (0.59–0.70)[a]
Long-term	LR-L1	MB-LR	ROS	0.95 (0.94–0.96)	0.86 (0.80–0.90)
	LR-L2	MB-LR	None	0.96 (0.95–0.97)	0.86 (0.81–0.90)
	LR-EN	MB-LR	SMOTE + ENN	0.92 (0.90–0.93)	0.83 (0.76–0.88)
	kNN	UFS-MI	TL	0.88 (0.86–0.90)	0.74 (0.68–0.80)
	SVM	RFE-LR	ENN	0.94 (0.92–0.96)	0.79 (0.73–0.85)
	ET	MB-LR	ENN	0.93 (0.92–0.94)	0.88 (0.84–0.91)[a]
	GTB	UFS-F	ROS	0.89 (0.86–0.91)	0.77 (0.71–0.83)
Longitudinal	LR-L1	UFS-MI	None	0.63 (0.57–0.68)	0.52 (0.41–0.61)
	LR-L2	RFE-LR	NCL	0.60 (0.55–0.66)	0.39 (0.29–0.48)
	LR-EN	UFS-MI	TL	0.62 (0.57–0.68)	0.52 (0.42–0.60)
	kNN	UFS-MI	NCL	0.65 (0.61–0.69)	0.58 (0.49–0.66)
	SVM	UFS-MI	OSS	0.66 (0.60–0.71)	0.57 (0.46–0.66)
	ET	UFS-MI	TL	0.66 (0.61–0.71)	0.51 (0.40–0.60)
	GTB	RFE-LR	ROS	0.68 (0.62–0.72)	0.63 (0.52–0.71)[a]

[a] *Best performing models at a given end point.*

Holm-Bonferroni correction, all the Friedman tests were significant at the FWER ≤ 0.05. Therefore, classification, feature selection, and sampling algorithms were compared for both the time-specific and the longitudinal models.

In the time-specific models, the support vector machine was by far the best scoring classifier, outperforming the other classifiers in over 70% of cases (**Figure 6**), whereas gradient tree boosting was on average the worst performing classifier (**Figure 7**). Conversely, gradient tree boosting together with support vector machines and extra-trees predicted xerostomia significantly better than all the other classifiers in the longitudinal approach.

The logistic regression-based algorithms performed significantly better than the feature selection methods based on extra-trees, in both the time-specific and the longitudinal models. Interestingly, while univariate feature selection by mutual information was the worst performing feature selection method in the time-specific models, it was one of the best in the longitudinal approach. Not performing feature selection was not disadvantageous in terms of predictive performance.

In both the time-specific and the longitudinal approach, no sampling algorithm gave a significant advantage over no sampling at all. In the time-specific models, Tomek links and the neighborhood cleaning rule performed significantly better than any oversampling algorithm. In the longitudinal models, Tomek links performed significantly better than random oversampling or ADASYN.

3.5. Generalization Performance

The best performing models stratified by end point and classifier are listed in **Table 4**. These models were retested by nested cross-validation to estimate their generalization performance. Early xerostomia (0–6 months after treatment) was predicted fairly well only by the k-nearest neighbors classifier (AUC = 0.65). The models of late xerostomia (6–15 months after treatment) generalized slightly better with logistic regression, k-nearest neighbors, and gradient tree boosting scoring AUC > 0.60. For long-term xerostomia (15–24 months after treatment), the models generalized best with the AUC ranging from 0.74 (k-nearest neighbors) to 0.88 (extra-trees). The longitudinal models failed to generalize except the gradient tree boosting classifier, which achieved AUC = 0.63. Generalization AUCs were on average 0.10 lower than tuning AUCs for all the analyzed end points.

3.6. Model Interpretation

Only the models predicting long-term xerostomia achieved high generalization scores, that is AUC > 0.70. For that reason, model interpretation was performed only for this end point. The multivariate models of long-term xerostomia relied mostly on the parotid gland volume, the spread of the contralateral dose–volume histogram, and the parotid gland eccentricity (**Figure 8**). The contralateral dose gradient in the right–left direction, despite good univariate predictive power, was included in only one model.

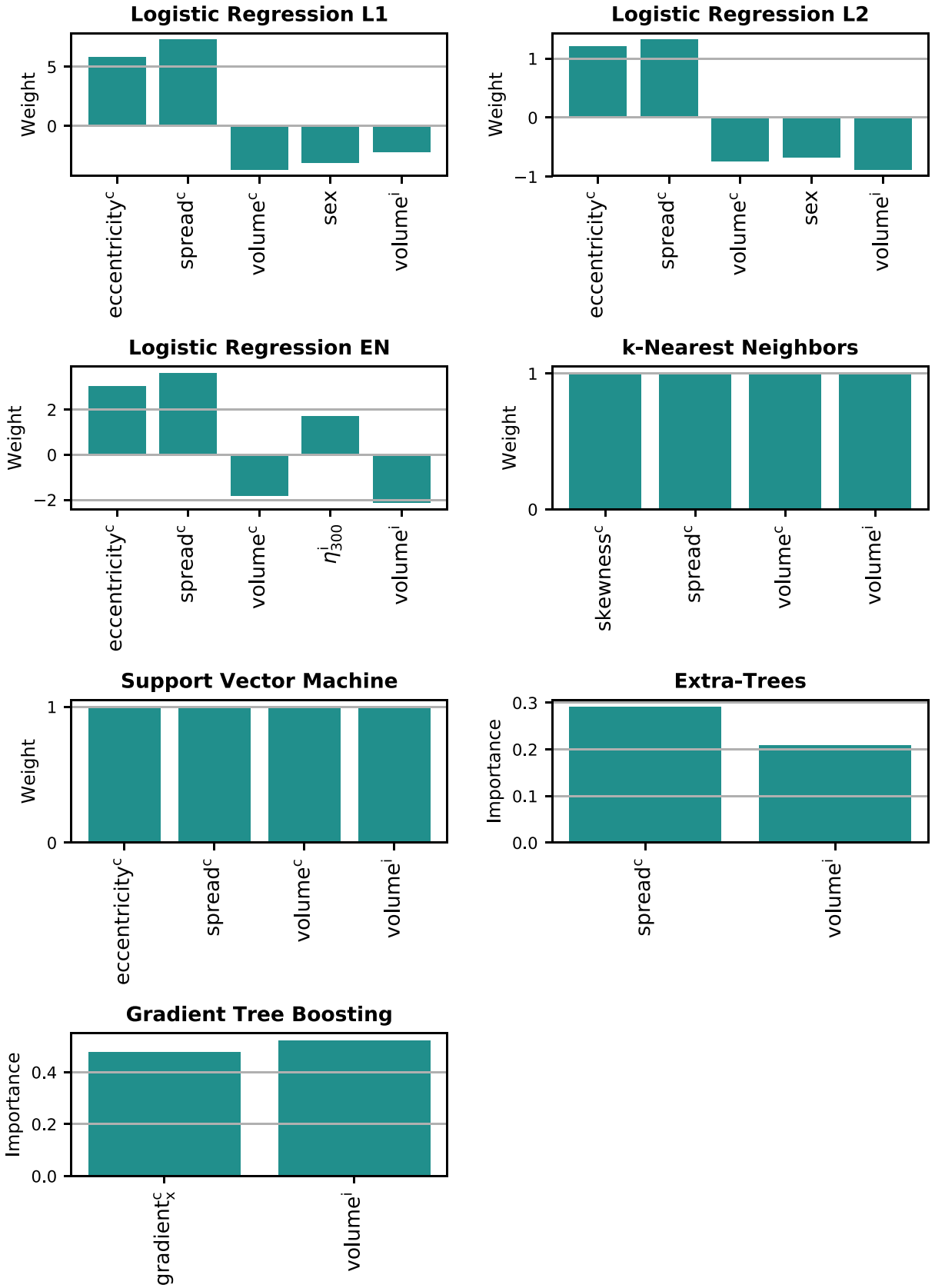

FIGURE 8 | Features underlying the multivariate models of long-term xerostomia. i, ipsilateral gland; c, contralateral gland.

4. DISCUSSION

The univariate analysis showed that parotid- and dose-shape features can be highly predictive of xerostomia. Patients with small parotid glands (median parotid volume in the positive group 9,557 vs. 14,374 mm^3 in the negative group) and steep dose gradients in the patient's right–left direction (median gradient in the positive group 1.7 vs. 1.2 Gy/mm in the negative group) were significantly more likely to develop long-term xerostomia. A possible explanation of this finding could be the fact that parotid glands typically shrink and move toward the medial direction during the course of radiotherapy. As a result, for patients with small parotid glands, the gradient is a proxy for the change of any dose-related metric subject to motion. As such, this might be an indicator of neglected motion and deformation effects during the modeling process.

Nevertheless, good discriminative power of the dose gradients and poor performance of the mean dose should be put into perspective of the previous studies validating mean-dose models. In cohorts where patients received a high radiation dose to parotid glands, the mean dose allowed achieving AUC above 0.80 (2, 3). It seems that inclusion of patients with less conformal treatment plans and a higher dosage to parotids would result in a cluster of patients with complications in the high-dose region of **Figure 4**. Therefore, for relatively high doses, the mean dose alone is a good xerostomia predictor irrespective of the dose gradient, whereas in the low-dose regime of modern radiotherapy treatments dose gradients are more informative and the mean dose is less predictive.

In the multivariate analysis, we did not find a model that would achieve generalization AUC above 0.65 for early or late-effects, even though a few univariate models of late xerostomia exceeded that value. Similarly, the multivariate models of long-term xerostomia, despite their good generalization scores ($AUC_{max} = 0.88$), performed on a par with the univariate models based on the parotid volume or the contralateral dose gradient in the patient's right–left direction. Comparable performance of the univariate and the multivariate models could be caused by the small sample size, especially the small minority class. In such setting, the distribution of model covariates can nonnegligibly differ between training and testing folds, hindering model training and reducing performance of the model.

The analysis of the multivariate models highlighted the importance of personalized treatment planning in radiotherapy. The models were strongly based on patient-specific and dose-independent features, such as parotid volume, parotid eccentricity, and the patient's sex. Females with small, elongated parotid glands were at higher risk of long-term xerostomia than males with large and rather round parotids. Interestingly, the dose gradient, despite relatively high predictive power, was included in only one model. Instead, the most common dosiomic feature was the spread of the contralateral dose–volume histogram quantifying the SD of the dose within a parotid gland. Nevertheless, due to the geometry of the problem, the DVH spread and spatial dose gradients measured a similar characteristic of the dose distribution. That is, a large spread of the DVH was present when part of the parotid gland received high dose, whereas another part was spared.

In the time-specific models, the support vector machine was most commonly the best classifier. The other classifiers performed similarly to one another. The unexceptional performance of the ensemble methods (extra-trees and gradient tree boosting) could stem from the fact that complex models need more training samples to correctly learn the decision boundary. Among the longitudinal models, we saw a more commonly observed classifier "ranking," that is GTB > ET > SVM > LR > kNN (19). Feature selection did not give a clear advantage over no feature selection in terms of the predictive performance. Nonetheless, feature selection allowed for a reduction of model complexity and made model interpretation easier. The best results were achieved with the logistic regression-based algorithms and feature selection by mutual information (only in the longitudinal models). We have not found evidence that sampling methods improve accuracy of predictions. Moreover, we observed that certain kinds of sampling, especially random oversampling, can significantly decrease predictive performance of the models.

Nested cross-validation proved to be an important step in the analysis. On average, the generalization AUCs were significantly lower than the AUCs achieved in model tuning. Our findings confirm the notion that single cross-validation can lead to overoptimistic performance estimates when hyperparameter tuning is involved in model building.

5. CONCLUSION

We demonstrated that in a highly conformal regime of modern radiotherapy, use of organ- and dose-shape features can be advantageous for modeling of treatment outcomes. Moreover, due to strong dependence on patient-specific factors, such as the parotid shape or the patient's sex, our results highlight the need for development of personalized data-driven risk profiles in future NTCP models of xerostomia.

Our results show that the choice of a classifier and a feature selection algorithm can significantly influence predictive performance of the NTCP model. Moreover, in relatively small clinical data sets, simple logistic regression can perform as well as top-ranking machine learning algorithms, such as extra-trees or support vector machines. We saw no significant advantage in using data cleaning or reducing the class imbalance. Our study confirms the need for significantly larger patient cohorts to benefit from advanced classification methods, such as gradient tree boosting. We showed that single cross-validation can lead to overoptimistic performance estimates when hyperparameter optimization is involved; either nested cross-validation or an independent test set should be used to estimate the generalization performance of a model.

LIST OF NON-STANDARD ABBREVIATIONS

Classification	
LR-L1	Logistic regression with L1 penalty
LR-L2	Logistic regression with L2 penalty
LR-EN	Logistic regression with elastic net penalty

(Continued)

kNN	k-Nearest neighbors
SVM	Support vector machine
ET	Extra-trees
GTB	Gradient tree boosting
Feature selection	
UFS-F	Univariate feature selection by F-score
UFS-MI	Univariate feature selection by mutual information
RFE-LR	Recursive feature elimination by logistic regression
RFE-ET	Recursive feature elimination by extra-trees
MB-LR	Model-based feature selection by logistic regression
MB-ET	Model-based feature selection by extra-trees
Sampling	
ROS	Random oversampling
SMOTE	Synthetic minority oversampling
ADASYN	Adaptive synthetic sampling
OSS	One-sided selection
TL	Tomek links
ENN	Wilson's edited nearest neighbor rule
NCL	Neighborhood cleaning rule
SMOTE + ENN	SMOTE followed by the ENN
SMOTE + TL	SMOTE followed by TL

ETHICS STATEMENT

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Heidelberg University. Nr. S-392/2016 "Validation and development of probabilistic prediction models for radiation-induced xerostomia."

AUTHOR CONTRIBUTIONS

HG, FS, HH, and MB contributed to the acquisition of the clinical data. HG, FS, and MB contributed to the analysis of the follow-up data. HG, FB, and MB contributed to the methodology. HG performed feature extraction, data visualization, statistical analysis, and drafted the manuscript. MB was the senior author supervising the project.

ACKNOWLEDGMENTS

We would like to thank (in alphabetical order) Jürgen Debus, Alexander Emig, Sebastian Klüter, Henning Mescher, Dieter Ötzel, and Kai Schubert for support during the extraction of treatment and patient data.

REFERENCES

1. Deasy JO, Moiseenko V, Marks L, Chao KSC, Nam J, Eisbruch A. Radiotherapy dose-volume effects on salivary gland function. *Int J Radiat Oncol Biol Phys* (2010) 76(3 Suppl):58–63. doi:10.1016/j.ijrobp.2009.06.090
2. Houweling AC, Philippens MEP, Dijkema T, Roesink JM, Terhaard CHJ, Schilstra C, et al. A comparison of dose-response models for the parotid gland in a large group of head-and-neck cancer patients. *Int J Radiat Oncol Biol Phys* (2010) 76(4):1259–65. doi:10.1016/j.ijrobp.2009.07.1685
3. Beetz I, Schilstra C, Burlage FR, Koken PW, Doornaert P, Bijl HP, et al. Development of NTCP models for head and neck cancer patients treated with three-dimensional conformal radiotherapy for xerostomia and sticky saliva: the role of dosimetric and clinical factors. *Radiother Oncol* (2012) 105(1):86–93. doi:10.1016/j.radonc.2011.05.010
4. Buettner F, Miah AB, Gulliford SL, Hall E, Harrington KJ, Webb S, et al. Novel approaches to improve the therapeutic index of head and neck radiotherapy: an analysis of data from the PARSPORT randomised phase III trial. *Radiother Oncol* (2012) 103(1):82–7. doi:10.1016/j.radonc.2012.02.006
5. Lee T-F, Liou M-H, Ting H-M, Chang L, Lee H-Y, Wan Leung S, et al. Patient- and therapy-related factors associated with the incidence of xerostomia in nasopharyngeal carcinoma patients receiving parotid-sparing helical tomotherapy. *Sci Rep* (2015) 5:13165. doi:10.1038/srep13165
6. Gabrys HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Parotid gland mean dose as a xerostomia predictor in low-dose domains. *Acta Oncol* (2017) 56(9):1197–203. doi:10.1080/0284186X.2017.1324209
7. Eisbruch A, Kim HM, Terrell JE, Marsh LH, Dawson LA, Ship JA. Xerostomia and its predictors following parotid-sparing irradiation of head-and-neck cancer. *Int J Radiat Oncol Biol Phys* (2001) 50(3):695–704. doi:10.1016/S0360-3016(01)01512-7
8. Lee T-F, Chao PJ, Ting HM, Chang L, Huang YJ, Wu JM, et al. Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer. *PLoS One* (2014) 9(2):e89700. doi:10.1371/journal.pone.0089700
9. Hawkins PG, Lee JY, Mao Y, Li P, Green M, Worden FP, et al. Sparing all salivary glands with IMRT for head and neck cancer: longitudinal study of patient-reported xerostomia and head-and-neck quality of life. *Radiother Oncol* (2018) 126(1):68–74. doi:10.1016/j.radonc.2017.08.002
10. Luijk PV, Pringle S, Deasy JO, Moiseenko VV, Faber H, Hovan A, et al. Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer. *Sci Transl Med* (2015) 7(305):1–8. doi:10.1126/scitranslmed.aac4441
11. van Dijk LV, Brouwer CL, van der Schaaf A, Burgerhof JGM, Beukinga RJ, Langendijk JA, et al. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. *Radiother Oncol* (2017) 122(2):185–91. doi:10.1016/j.radonc.2016.07.007
12. van Dijk LV, Brouwer CL, Paul H, Laan VD, Johannes GM, Langendijk JA, et al. Geometric image biomarker changes of the parotid gland are associated with late xerostomia. *Int J Radiat Oncol Biol Phys* (2017) 99(5):1101–10. doi:10.1016/j.ijrobp.2017.08.003
13. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. *Phys Med Biol* (2009) 54(18):S9–30. doi:10.1088/0031-9155/54/18/S02
14. Gulliford S. Modelling of normal tissue complication probabilities (NTCP): review of application of machine learning in predicting NTCP. In: El Naqa I, Li R, Murphy MJ, editors. *Machine Learning in Radiation Oncology*. Cham: Springer (2015). p. 277–310.
15. Dean JA, Welsh LC, Wong KH, Aleksic A, Dunne E, Islam MR, et al. Normal tissue complication probability (NTCP) modelling of severe acute mucositis using a novel oral mucosal surface organ at risk. *Clin Oncol* (2017) 29(4):263–73. doi:10.1016/j.clon.2016.12.001
16. Chen S, Zhou S, Yin F-F, Marks LB, Das SK. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. *Med Phys* (2007) 34(10):3808–14. doi:10.1118/1.2776669
17. Ospina JD, Zhu J, Chira C, Bossi A, Delobel JB, Beckendorf V, et al. Random forests to predict rectal toxicity following prostate cancer radiation therapy. *Int J Radiat Oncol Biol Phys* (2014) 89(5):1024–31. doi:10.1016/j.ijrobp.2014.04.027
18. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* (2005) 21(5):631–43. doi:10.1093/bioinformatics/bti033
19. Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. *Data-Driven Advice for Applying Machine Learning to Bioinformatics Problems*. (2017). *ArXiv*.
20. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine learning classifiers for prognostic biomarkers of head & neck cancer. *Front Oncol* (2015) 5:272. doi:10.3389/fonc.2015.00272
21. National Cancer Institute (U.S.). *Common Terminology Criteria for Adverse Events (CTCAE) v4.03*. Bethesda, MD: U.S. Department of Health and Human Services (2010).

22. Salkind NJ. *Encyclopedia of Measurement and Statistics*. Thousand Oaks: SAGE Publications (2007). p. 508–10.
23. Eisbruch A, Ten Haken RK, Kim HM, Marsh LH, Ship JA. Dose, volume, and function relationships in parotid salivary glands following conformal and intensity-modulated irradiation of head and neck cancer. *Int J Radiat Oncol Biol Phys* (1999) 45(3):577–87. doi:10.1016/S0360-3016(99)90269-9
24. Roesink JM, Moerland MA, Battermann JJ, Hordijk GJ, Terhaard CH. Quantitative dose-volume response analysis of changes in parotid gland function after radiotherapy in the head-and-neck region. *Int J Radiat Oncol Biol Phys* (2001) 51(4):938–46. doi:10.1016/S0360-3016(01)01717-5
25. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver characteristic (ROC) curve. *Radiology* (1982) 143:29–36. doi:10.1148/radiology.143.1.7063747
26. Qin G, Hotilovac L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat Methods Med Res* (2008) 17(2):207–21. doi:10.1177/0962280207087173
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* (1995) 57(1):289–300.
28. Gavrilov Y, Benjamini Y, Sarkar SK. An adaptive step-down procedure with proven FDR control under independence. *Ann Stat* (2009) 37(2):619–29. doi:10.1214/07-AOS586
29. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *Intell Data Anal* (2002) 6(5):429–49.
30. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* (2009) 21(9):1263–84. doi:10.1109/TKDE.2008.239
31. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* (2003) 3:1157–82.
32. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* (2012) 13:281–305.
33. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* (2005) 21(15):3301–7. doi:10.1093/bioinformatics/bti499
34. Krzanowski W, Hand D. Assessing error rate estimators: the leave-one-out method reconsidered. *Aust N Z J Stat* (1997) 39(1):35–46. doi:10.1111/j.1467-842X.1997.tb00521.x
35. Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput Stat Data Anal* (2011) 55(4):1828–44. doi:10.1016/j.csda.2010.11.018
36. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* (1979) 6:65–70.
37. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* (2010) 11:2079–107.
38. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res* (2017) 18(17):1–5.
39. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* (2007) 9(3):99–104. doi:10.1109/MCSE.2007.55
40. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* (2011) 13(2):22–30. doi:10.1109/MCSE.2011.37
41. Demšar J, Curk T, Erjavec A, Hočevar T, Milutinovič M, Možina M, et al. Orange: data mining toolbox in Python. *J Mach Learn Res* (2013) 14:2349–53.
42. McKinney W. Data structures for statistical computing in Python. In: van der Walt S, Millman J, editors. *SciPy 2010: Proceedings of the 9th Python in Science Conference*. Austin, TX, USA. (2001) p. 51–6.
43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
44. Chen T, Guestrin C. *XGBoost: A Scalable Tree Boosting System*. (2016). p. 1–6. arXiv Prepr. arXiv1603.02754v3.
45. Gonzalez RC, Woods RE. *Digital Image Processing*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, Inc (2006).
46. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* (2002) 16:321–57.
47. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In *Proc 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, China (2008). p. 1322–8.
48. Tomek I. Two modifications of CNN. *IEEE Trans Syst Man Cybern* (1976) 6:769–72.
49. Hart PE. The condensed nearest neighbour rule. *IEEE Trans Inf Theory* (1968) 14(5):515–6. doi:10.1109/TIT.1968.1054155
50. Kubat M, Matwin S. Addressing the course of imbalanced training sets: one-sided selection. In: Fisher DH, editor. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*. Nashville, TN, USA/San Francisco: Morgan Kaufmann (1997). p. 179–86.
51. Wilson DR. Asymptotic properties of nearest neighbor rules using edited data. *Inst Electr Electron Eng Trans Syst Man Cybern* (1972) 2(3):408–21.
52. Laurikkala J. Improving identification of difficult small classes by balancing class distribution. In: Quaglini S, Barahona P, Andreassen S, editors. *AIME 2001 Artificial Intelligence in Medicine: Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe*. Cascais, Portugal/Berlin: Springer (2001) p. 63–6.
53. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explor Newsl* (2004) 6(1):20–9. doi:10.1145/1007730.1007735
54. Gu Q, Li Z, Han J. Generalized Fisher Score for feature selection. *CoRR* (2012) 3:327–30.
55. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. In: Aggarwal CC, editor. *Data Classification Algorithms and Applications*. Boca Raton, FL: CRC Press (2014). p. 37–64.
56. Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York, NY: John Wiley and Sons (2012).
57. Lowry R, editor. One-way analysis of variance for independent samples. *Concepts and Applications of Inferential Statistics*. Poughkeepsie, NY: DOER – Directory of Open Educational Resources (2014).
58. Murphy KP. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press (2012).
59. Kohavi R, John G. Wrappers for feature subset selection. *Artif Intell* (1997) 97(97):273–324. doi:10.1016/S0004-3702(97)00043-X
60. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn* (2002) 46(1–3):389–422. doi:10.1023/A:1012487302797
61. Hastie T, Tibshirani RJ, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2 ed. New York, NY: Springer (2009).
62. Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: Brodley C, editor. *ICML 2004: Proceedings of the Twenty-First International Conference on Machine Learning*. Banff, Alberta, Canada/New York: ACM (2004). 78 p.
63. Bishop CM. *Pattern Recognition and Machine Learning*. 1 ed. New York, NY: Springer (2006).
64. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* (2005) 67:301–20. doi:10.1111/j.1467-9868.2005.00527.x
65. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* (1998) 2:121–67. doi:10.1023/A:1009715923555
66. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn* (2006) 63(1):3–42. doi:10.1007/s10994-006-6226-1
67. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* (1997) 55(1):119–39. doi:10.1006/jcss.1997.1504

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

The MATLAB code used for feature extraction is available on GitHub https://github.com/hubertgabrys/DicomToolboxMatlab.

A. Parotid Shape

A.1. Volume

Volume V of the parotid gland.

A.2. Surface area

Surface area A of the parotid gland.

A.3. Sphericity

Parotid gland sphericity was defined as the ratio of the surface area of a sphere of the same volume as the parotid gland to the actual surface area of the parotid

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}.$$

A.4. Compactness

Parotid gland compactness was defined as a ratio of the parotid gland surface area to the parotid gland volume.

$$\kappa = \frac{A}{V}.$$

A.5. Eccentricity

Eccentricity ε measured how elongated the parotid gland was. Larger asymmetry of the gland corresponded to larger values of ε.

$$\varepsilon = 1 - \sqrt{\frac{\lambda_{min}}{\lambda_{max}}},$$

where eigenvalues λ_i of the parotid shape covariance matrix correspond to the dimensions of the parotid gland along the principal axes defined by the eigenvectors. The covariance matrix is defined as:

$$\mathrm{Cov}[I(x,y,z)] = \begin{pmatrix} \mu_{200} & \mu_{110} & \mu_{101} \\ \mu_{110} & \mu_{020} & \mu_{011} \\ \mu_{101} & \mu_{011} & \mu_{002} \end{pmatrix}$$

$$\mu_{pqr} = \sum_{x,y,z} (x-\bar{x})^p (y-\bar{y})^q (z-\bar{z})^r I(x,y,z),$$

$$\bar{x} = \frac{\sum_{x,y,z} xI(x,y,z)}{\sum_{x,y,z} I(x,y,z)}$$

where x, y, and z are the coordinates of the voxel, $I(x,y,z)$ the indicator function indicating whether a voxel belongs to the parotid, and μ_{pqr} central moments of the parotid. $\bar{y}$ and $\bar{z}$ were defined analogously to $\bar{x}$.

B. Dose–Volume Histogram

B.1. Mean

The mean dose to the parotid gland.

B.2. Spread

The spread of the differential dose–volume histogram was quantified by the SD of the dose within the parotid gland.

B.3. Skewness

The skewness of the differential dose–volume histogram was measured by the third standardized moment. Negative skewness corresponds to the dose–volume histogram skewed toward lower dose, whereas positive skewness means the dose–volume histogram is skewed toward higher dose.

B.4. Dx

The minimum dose to x% "hottest" volume of the parotid gland.

B.5. Vx

Percentage volume of the parotid gland receiving at least x Gy.

B.6. Entropy

Entropy H measures smoothness of the dose within the parotid gland (45):

$$H = -\sum_{i=1}^{256} m(d_i) \log m(d_i),$$

where d_i is the dose delivered to the ith voxel and $m(d_i)$ is the corresponding histogram. $H = 0$ for a uniform dose and $H > 0$ for a nonuniform dose.

B.7. Uniformity

Uniformity U of the dose within the parotid gland (45):

$$U = \sum_{i=1}^{256} m^2(d_i),$$

$U = 1$ for a uniform dose and $U < 1$ for a nonuniform dose.

C. Subvolume Mean Dose

Parotid gland subvolumes were defined by axial, coronal, and sagittal slices that cut parotid glands in thirds along the patient's axes. The cuts were positioned in such a way that each subvolume comprised approximately the same number of voxels. As a result, nine, not exclusive, subvolumes were defined: three in x, three in y, and three in z direction. For each subvolume the mean radiation dose was calculated, e.g., the mean dose to the anterior third of the parotid gland (s_y^1) or the mean dose to the superior third of the parotid gland (s_z^3).

D. Dose Gradients

Average dose gradients measured average change of the dose along one of patient axes and were defined as:

$$\mathrm{Gradient_x} = \frac{\sum_{x,y,z} D(x+1,y,z)I(x+1,y,z) - D(x-1,y,z)I(x-1,y,z)}{2\sum_{x,y,z} I(x,y,z)},$$

where x, y, and z are the coordinates of the voxel, $D(x,y,z)$ the dose delivered to the voxel, and $I(x,y,z)$ the indicator function indicating whether a voxel belongs to the parotid. $\mathrm{Gradient_y}$ and $\mathrm{gradient_z}$ were defined analogously to $\mathrm{gradient_x}$.

E. Three-Dimensional Dose Moments

The scale invariant dose moments allowed to quantify three-dimensional shape of the dose distribution within the parotid gland. Visualization of the moments can be found in Buettner et al. Supplementary Figure 1–3 (4). The moments were defined as:

$$\eta_{pqr} = \frac{\sum_{x,y,z} (x-\bar{x})^p (y-\bar{y})^q (z-\bar{z})^r D(x,y,z) I(x,y,z)}{\left(\sum_{x,y,z} D(x,y,z) I(x,y,z)\right)^{\frac{p+q+r}{3}+1}}$$

$$\bar{x} = \frac{\sum_{x,y,z} x I(x,y,z) D(x,y,z)}{\sum_{x,y,z} I(x,y,z) D(x,y,z)},$$

$\bar{y}$ and $\bar{z}$ were defined analogously. In particular, we considered moments quantifying dose variance, covariance, skewness, and coskewness.

E.1. Dose Variance (η_{200}, η_{020}, η_{002})

Dose variance corresponds to the spread of the dose along a given direction.

E.2. Dose Covariance (η_{110}, η_{101}, η_{011})

Dose covariance measures how the dose covaries along two axes. For example, positive values of η_{110} correspond to dose deposition along *xy* direction, whereas negative values correspond to dose deposition along the direction perpendicular to *xy*.

E.3. Dose Skewness (η_{300}, η_{030}, η_{003})

Dose skewness measures asymmetry of the dose distribution along a given axis.

E.4. Dose Coskewness (η_{210}, η_{201}, η_{120}, η_{021}, η_{012}, η_{102})

Dose coskewness measures how dose variance along one direction covaries with another dimension, e.g., negative value of η_{210} would mean that variance of the dose along *x* axis increases when moving up the *y* axis.

APPENDIX B

It has been reported that class imbalance together with low size of the minority class can hinder the performance of predictive models. There are two approaches commonly taken to alleviate this problem: oversampling and undersampling. In oversampling, one reduces the imbalance between classes by random replication or synthetic creation of minority class observations. Conversely, in undersampling the majority class size is reduced by elimination of its observations. Additionally, there are data cleaning methods which, through undersampling, aim to remove the observations that are considered noise or the observations close to the decision boundary, irrespective of their class membership. As a result, data cleaning methods do not reduce class imbalance but rather improve definitions of class clusters. Hyperparameters used to tune the sampling and the data cleaning algorithms are listed in **Table A1**.

A. Random Oversampling

The data set imbalance is reduced by randomly duplicating observations from the minority class.

TABLE A1 | Hyperparameters used to tune the sampling algorithms.

Algorithm	Hyperparameters	Values
ROS	–	–
SMOTE	**k_neighbors**: Number of nearest neighbors used to construct synthetic samples.	{3,4,5}
	m_neighbors: Number of nearest neighbors used to determine if a minority sample is in danger.	{7,8,9}
	kind: Type of SMOTE algorithm.	{"regular," "borderline1," "borderline2"}
ADASYN	**n_neighbors**: Number of nearest neighbors to use to construct synthetic samples.	{3,5,8}
OSS	–	–
TL	–	–
ENN	**n_neighbors**: Number of nearest neighbors.	{2,3,5}
	kind_sel: Type of ENN algorithm.	{"all," "mode"}
NCL	**n_neighbors**: Number of nearest neighbors.	{2,3,5}
SMOTE + TL	–	–
SMOTE + ENN	–	–

Hyperparameters not listed in this table assumed the default values of imbalanced-learn package (38).

B. Synthetic Minority Oversampling

Synthetic minority oversampling (SMOTE) was proposed by Chawla et al. (46). The algorithm generates new synthetic minority observations by considering *k* nearest neighbors of a randomly selected minority observation. Next, the difference between the observation feature vector and one of the nearest neighbors feature vector is taken. This difference is then multiplied by a random weight between 0 and 1, and added to the observation feature vector to generate a new synthetic observation. In SMOTE, approximately equal number of synthetic observations is created for each minority class observation.

C. Adaptive Synthetic Sampling

Adaptive synthetic sampling (ADASYN) (47), similarly to SMOTE, generates synthetic minority class observations by interpolating feature vectors between a minority class observation and a randomly selected nearest neighbor. The key difference to SMOTE is that ADASYN aims to create more synthetic data for minority class observations that are hard to learn. For that reason, a learning difficulty weight is calculated for each minority class observation, based on the number of majority class observations in its neighborhood. Based on these weights, more synthetic observations are created for "difficult" minority class observations.

D. Tomek Links

A pair of observations (E_i,E_j) stemming from different classes and with distance $d(E_i, E_j)$ form a Tomek link if there is no observation E_l, such that $d(E_i, E_l) < d(E_i, E_j)$ or $d(E_j, E_l) < d(E_i, E_j)$ (48). As an undersampling method, all the observations in the majority class forming Tomek links are removed; when used as a data

cleaning method, both the observation from the majority and the observation from the minority class are eliminated.

E. Condensed Nearest Neighbor Rule

The condensed nearest neighbor rule (CNN) proposed by Hart (49) undersamples the data set to find a consistent subset $\hat{E}$ of all observations E. First, all minority class observations and one randomly selected majority class observation are moved to $\hat{E}$. Next, the rest of the majority class observations are classified using 1-nearest neighbor rule and during this process every misclassified observation is moved to subset $\hat{E}$. The procedure continues until all misclassified observations are in the subset $\hat{E}$ (50). Intuitively, CNN reduces the number of redundant observations in majority class that are far from the decision border and therefore less informative in learning.

F. One-Sided Selection

One-sided selection (OSS) (50) is an undersampling method realized by Tomek links algorithm followed by CNN. Tomek links undersample the majority class and remove noisy and borderline class observations. CNN, on the other hand, removes observations from the majority class that are distant from the decision border and likely are not informative.

G. Wilson's Edited Nearest Neighbor Rule

The Wilson's edited nearest neighbor rule (ENN) (51) removes all observations which class label differ from the class of its k nearest neighbors.

H. Neighborhood Cleaning Rule

The neighborhood cleaning rule (NCL) (52) is a modification of the ENN algorithm. As in the ENN, the class of each observation is compared with the classes of its k nearest neighbors. If the analyzed observation belongs to the majority class, the procedure is the same as in the ENN. However, if the observation belongs to the minority class and its k nearest neighbors to the majority class, the minority class observation is kept in the data set and the k nearest neighbors are removed.

I. SMOTE + TL

First, the original data set is oversampled with SMOTE, and then Tomek links are identified and removed. The method aims to produce a balanced data set with well-defined class clusters (53).

J. SMOTE + ENN

This method is similar to SMOTE + TL but with stronger data cleaning component realized by the ENN (53).

APPENDIX C

Feature selection is a crucial part of model building. It not only allows to improve accuracy of model predictions but also reduces the dimensionality of the input space. A reduced dimensionality of the input space decreases the risk of model overfitting and improves model interpretability. Hyperparameters used to tune the feature selection algorithms are listed in **Table A2**.

TABLE A2 | Hyperparameters used to tune the feature selection algorithms.

Algorithm	Hyperparameters	Values
UFS-F	**k**: Number of features to select.	{2,3,4,5,6}
UFS-MI	**k**: Number of features to select.	{2,3,4,5,6}
RFE-LR	**k**: Number of features to select.	{2,3,4,5,6}
	step: Number of features to remove at each iteration.	1
	class_weight: Whether class weights are equal or inversely proportional to class frequencies.	{None, "balanced"}
	C: Inverse of regularization strength.	$\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$
	penalty: Type of regularization.	"l2"
RFE-ET	**k**: Number of features to select.	{2,3,4,5,6}
	step: Fraction of features to remove at each iteration.	0.5
	class_weight: Whether class weights are equal or inversely proportional to class frequencies.	{None, "balanced," "balanced_subsample"}
	n_estimators: Number of decision trees.	[90,140]
MB-LR	**k**: Number of features to select.	{2,3,4,5,6}
	class_weight: Whether class weights are equal or inversely proportional to class frequencies.	{None, "balanced"}
	C: Inverse of regularization strength.	$\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$
	penalty: Type of regularization.	{"l1," "l2"}
MB-ET	**k**: Number of features to select.	{2,3,4,5,6}
	class_weight: Whether class weights are equal or inversely proportional to class frequencies.	{None, "balanced," "balanced_subsample"}
	n_estimators: Number of decision trees.	[90,140]

Hyperparameters not listed in this table assumed the default values of scikit-learn package (43).

A. Univariate Feature Selection

Univariate feature selection methods evaluate each feature separately relying solely on the relation between one feature characteristic and the modeled variable. After all the features were graded, the features with the highest rankings are selected. A disadvantage of univariate feature selection is that the algorithm fails to select features which have relatively low individual scores but a high score when combined together. Also, due to the fact that univariate feature selection methods evaluate features individually, they are unable to handle feature redundancy (54, 55).

A.1. Fisher Score

Intuitively, Fisher score is a ratio of the between-class scatter to the within-class scatter. As a result, high Fisher scores correspond to features with well defined class clusters (low within-class scatter) that are distant from each other (large between-class scatter) (56). Fisher score is commonly used in supervised classification tasks due to its low computational cost and general good performance (54).

Fisher score of feature X was calculated using the following formula (57):

$$F(X) = \frac{\frac{1}{C-1}\sum_{c=1}^{C} N_c(\bar{x}_c - \bar{x})^2}{\frac{1}{N-C}\sum_{c=1}^{C}\sum_{i:y_i=c}(x_i - \bar{x}_c)^2}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{x}_c = \frac{1}{N_c}\sum_{i:y_i=c} x_i,$$

where C is the number of classes, N total number of observations, N_c number of observations in class c, $\bar{x}$ mean value of feature X, and $\bar{x}_c$ mean value of feature X in class c.

A.2. Mutual Information

This univariate feature selection method measures mutual information between each feature and the modeled variable. Intuitively, mutual information measures how much knowing the feature X value reduces uncertainty about the class label Y, and vice versa (58). This can be expressed by the formula:

$$MI(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

where $H(X)$ is the entropy of X and $H(X|Y)$ is the entropy of X after observing class Y.

$$H(X) = -\sum_{i=1}^{N} p(x_i)\log p(x_i)$$

$$H(X|Y) = -\sum_{i=1}^{N} p(y_i)\sum_{k=1}^{N} p(x_k|y_i)\log p(x_k|y_i).$$

Features with high mutual information are considered informative and are selected.

B. Recursive Feature Elimination

In the first step of recursive feature elimination (RFE), an induction algorithm is trained using the full set of features. Next, the features are ranked according to a given criterion, such as feature weight in logistic regression or feature importance in ensemble models. Then, the feature or the features with the smallest ranks are removed from the feature set. This procedure is repeated iteratively until the desired number of features is achieved (59, 60).

In contrast to univariate feature selection, recursive feature elimination methods can capture feature interactions. For that reason it can select not only good univariate predictors but also features which have low predictive power alone but high predictive power when pooled together.

The ability to handle feature redundancy depends on the induction algorithm used with RFE. For instance, L1-penalized logistic regression tends to select one of highly correlated features, hence reducing feature redundancy (61). On the contrary, L2-penalized logistic regression tends to give similar weights to correlated features, distributing the total feature importance among them. For the recursive feature elimination, we used two induction algorithms: logistic regression and extra-trees.

C. Model-Based Feature Selection

Model-based feature selection can be considered a special case of recursive feature elimination with only one iteration step. The induction algorithm is trained using the full set of features and the desired number of lowest scoring features is removed.

TABLE A3 | Hyperparameters used to tune the classification algorithms.

Algorithm	Hyperparameters	Values
LR-L1	**class_weight**: Whether class weights are equal or inversely proportional to class frequencies. **C**: Inverse of regularization strength.	{None, "balanced"} $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$
LR-L2	**class_weight**: Whether class weights are equal or inversely proportional to class frequencies. **C**: Inverse of regularization strength.	{None, "balanced"} $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$
LR-EN	**class_weight**: Whether class weights are equal or inversely proportional to class frequencies. **alpha**: Regularization strength. **l1_ratio**: Ratio between L1 and L2 penalty.	{None, "balanced"} $\{2^{-10}, 2^{-9.985}, 2^{-9.97}, \ldots, 2^{5}\}$ [0,1]
kNN	**n_neighbors**: Number of nearest neighbors. **p**: Power parameter of the Minkowski distance.	{1,2,3,. . .,9} {1,2,∞}
SVM	**class_weight**: Whether class weights are equal or inversely proportional to class frequencies. **C**: Inverse of regularization strength. **gamma**: Parameter of the RBF kernel.	{None, "balanced"} $\{2^{-5}, 2^{-4.985}, 2^{-4.97}, \ldots, 2^{10}\}$ $\{2^{-15}, 2^{-14.982}, 2^{-14.964}, \ldots, 2^{3}\}$
ET	**n_estimators**: Number of decision trees. **class_weight**: Whether class weights are equal of inversely proportional to class frequencies. **criterion**: The function to measure the quality of a split. **max_features**: Number of features to consider when calculating the best split. **min_samples_split**: The minimum number of samples required to split a node. **min_samples_leaf**: The minimum number of samples required to be at a leaf node.	[90, 230] {None, "balanced"} {"gini," "entropy"} {0.05, 0.10, 0.15,. . .,1} {2,3,4,. . .,20} {1,2,3,. . .,20}
GTB	**n_estimators**: Number of decision trees. **learning_rate**: Boosting learning rate. **max_depth**: Maximum tree depth. **gamma**: Minimum loss reduction required to make a further partition on a leaf node of the tree. **min_child_weight**: Minimum sum of instance weight(hessian) needed in a child. **subsample**: Ratio of the training samples used to grow trees. **reg_lambda**: L1 regularization term on weights. **reg_alpha**: L2 regularization term on weights.	[200, 2000] $\{2^{-7}, 2^{-6.994}, 2^{-6.988}, \ldots, 2^{-1}\}$ {1,2,3,. . .,6} {0.05,0.1,0.3,0.5,0.7,0.9,1} {1,3,5,7} {0.6,0.65,0.70,. . .,1} [0,1] [0,1]

Hyperparameters not listed in this table assumed the default values of scikit-learn (43) and xgboost (44) packages.

Similarly to RFE, we employed logistic regression and extra-trees as the induction algorithms.

APPENDIX D

The selection of the classifier is a critical part of model building, which directly determines the flexibility of the decision boundary. On the one hand, a too flexible model can result in overfitting and low generalizability. On the other hand, a too simple model can fail to capture the complexity of the true decision boundary and result in underfitting. Furthermore, the interpretability of the model depends strongly on the type of the chosen algorithm. Hyperparameters used to tune the classification algorithms are listed in **Table A3**.

A. Logistic Regression

Logistic regression is a simple linear model allowing to estimate probability of a binary response based on a number of risk factors. In order to avoid overfitting, logistic regression is usually regularized *via* L1, L2, or elastic net penalty. L1 penalty outperforms L2 penalty in terms of handling irrelevant and redundant features (62). Its ability to bring feature weights to zero results in sparse models and improves model interpretability (63). On the other hand, L1 tends to randomly select one of highly correlated features which can result in model variability (64). The elastic net method brings in a way the two worlds together and applies a penalty that is a convex combination of L1 and L2 regularization (64).

The advantages of logistic regression are its simplicity, interpretability, and easy tuning (only one hyperparameter with L1 or L2 regularization or two hyperparameters with elastic net regularization). The biggest disadvantage is a linear hypersurface decision boundary that may not be flexible enough to describe the real decision boundary.

B. k-Nearest Neighbors

The k-nearest neighbor (kNN) classifier looks at the k points in the training set that are nearest to the test input. The object is classified based on a majority vote of its neighbors (58). kNN has a much more flexible decision boundary compared to logistic regression. It will likely outperform logistic regression when the true decision boundary is highly irregular. Nevertheless, the curse of dimensionality has a considerable impact on the performance of the k-nearest neighbors classifier making feature selection crucial when working with high-dimensional data sets.

C. Support Vector Machine

Similarly to the k-nearest neighbors algorithm, the support vector machine does not learn a fixed set of parameters corresponding to the features of the input. It rather remembers the training examples and classifies new observations based on some similarity function. The two main concepts behind support vector machines are the kernel trick and the large margin principle. The kernel trick guarantees high flexibility of the decision boundary by allowing to operate in feature spaces of very high, even infinite, dimensionality. The large margin principle ensures model sparsity by discarding all observations not laying on maximum margin hypersurfaces. Support vector machines proved to be very successful in various classification tasks, including NTCP modeling. Unfortunately, interpretation of support vector machines with nonlinear kernels is a challenge (65).

D. Extra-Trees

The extra-trees classifier is an ensemble of decision trees. Each tree is built either on the full learning sample or on a bootstrap replica. At each node, a random subset of features is selected and for each feature a random cut-point is drawn. The best feature-cutpoint pair is selected to split the node. The tree is grown until the minimum sample size for splitting a node is reached. The ensemble predictions are the results of the majority vote of predictions of individual trees (66). A big advantage of the extra-trees algorithm is that it works "out-of-the-box" with no or minimal hyperparameter tuning.

E. Gradient Tree Boosting

Similarly to extra-trees, gradient tree boosting uses an ensemble of decision trees. Gradient tree boosting iteratively fits small decision trees to the data set in an adaptive fashion. After each iteration, training samples are reweighted to focus on the instances misclassified by the previous trees. When all trees are grown, the prediction is obtained by the weighted majority vote of the trees (61, 67).

Gradient tree boosting proved to be a very successful algorithm often outperforming neural networks, support vector machines, and other ensemble models. However, tuning the hyperparameters may be challenging.

frontiers in Oncology

ORIGINAL RESEARCH
published: 19 March 2018
doi: 10.3389/fonc.2018.00057

An Ensemble Approach to Knowledge-Based Intensity-Modulated Radiation Therapy Planning

*Jiahan Zhang[1], Q. Jackie Wu[1], Tianyi Xie[1], Yang Sheng[1], Fang-Fang Yin[1] and Yaorong Ge[2]**

[1] Department of Radiation Oncology, Duke University Medical Center, Durham, NC, United States, [2] Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, United States

Knowledge-based planning (KBP) utilizes experienced planners' knowledge embedded in prior plans to estimate optimal achievable dose volume histogram (DVH) of new cases. In the regression-based KBP framework, previously planned patients' anatomical features and DVHs are extracted, and prior knowledge is summarized as the regression coefficients that transform features to organ-at-risk DVH predictions. In our study, we find that in different settings, different regression methods work better. To improve the robustness of KBP models, we propose an ensemble method that combines the strengths of various linear regression models, including stepwise, lasso, elastic net, and ridge regression. In the ensemble approach, we first obtain individual model prediction metadata using in-training-set leave-one-out cross validation. A constrained optimization is subsequently performed to decide individual model weights. The metadata is also used to filter out impactful training set outliers. We evaluate our method on a fresh set of retrospectively retrieved anonymized prostate intensity-modulated radiation therapy (IMRT) cases and head and neck IMRT cases. The proposed approach is more robust against small training set size, wrongly labeled cases, and dosimetric inferior plans, compared with other individual models. In summary, we believe the improved robustness makes the proposed method more suitable for clinical settings than individual models.

Keywords: treatment planning, dose volume histogram prediction, regression model, machine learning, ensemble model, statistical modeling

OPEN ACCESS

Edited by:
Jun Deng,
Yale University,
United States

Reviewed by:
Sunyoung Jang,
Princeton Radiation Oncology,
United States
John C. Roeske,
Loyola University Medical Center, United States

****Correspondence:***
Yaorong Ge
yge@uncc.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *09 November 2017*
Accepted: *21 February 2018*
Published: *19 March 2018*

Citation:
Zhang J, Wu QJ, Xie T, Sheng Y, Yin F-F and Ge Y (2018) An Ensemble Approach to Knowledge-Based Intensity-Modulated Radiation Therapy Planning. Front. Oncol. 8:57. doi: 10.3389/fonc.2018.00057

INTRODUCTION

In radiation therapy, high quality treatment plans are crucial for reducing the possibility of normal tissue complications while maintaining good dose coverage of planning target volume (PTV). For intensity-modulated radiation therapy (IMRT), it is especially important to fully utilize the healthy tissue sparing potential enabled by the advanced treatment delivering system. However, the optimal achievable organ-at-risk (OAR) sparing is not known pre-planning, and planners need to rely on their previous experience, which makes the planning process subjective, iterative, and susceptible to intra- and inter-planner variation.

Knowledge-based planning (KBP) (1–5) has been shown to be a powerful tool for guiding planners and physicians to optimal achievable OAR dose volume histograms (DVHs) based on previous cases planned by experienced planners. In a previously proposed regression-based KBP framework (2), the workflow is as follows: (i) principle component analysis (PCA) is conducted for OAR DVHs in the training set, and the first three principle component scores (PCS) and corresponding basis

vectors are stored; (ii) pre-determined geometry information related to treatment planning goals, also referred to as features, are calculated for each patient; (iii) PCS of OAR DVH are fitted to features to generate a prediction model; (iv) features are calculated for new patients; and (v) best achievable OAR DVHs are calculated for new patients using the fitted model and the previously calculated PCA basis vectors.

In step (iii) of the previous framework, stepwise regression is used to select features and estimate the linear model. The method automatically picks several most important features step by step based on the significance of features. This approach is easy to implement and the output is interpretable. With careful training data preprocessing and feature selection, stepwise has achieved good results in OAR DVH prediction in research settings (6–12). However, there are some theoretical issues about this procedure, which could potentially result in some instabilities of the overall model training process. While stepwise regression has been very successful in the context of KBP, potential disadvantages of stepwise regression are well documented. First, it potentially suffers from overfitting if the size of the training set is relatively small compared to the number of features. This is because the procedure attempts to fit many models and the *p*-values, which are used as feature selection criteria, are not corrected for the number of hypothesis tested. In addition, stepwise regression does not cope with collinear features well. If two features are highly collinear, stepwise usually selects just one and discard the other. Ideally, if several collinear features are predictive of the outcome, all of these features should be selected to prevent overfitting and reduce model variance.

The purpose of this study is to improve the regression modeling aspect of KBP. Empirically, different regression methods perform well in different scenarios, such as different number of training cases, presence of collinear features, and presence of outlier cases. In this work, we develop an ensemble learning method to combine the strengths of these individual models and improve KBP model robustness.

MATERIALS AND METHODS

Individual Models

As a comparison to our proposed ensemble model, we study four individual regression models, including ridge regression (13, 14), lasso (15), elastic net (16), and stepwise regression with forward feature selection. These models also serve as base learners for the final ensemble model. The latter three models share the same objective function

$$\beta = \operatorname{argmin}\left\{\left\|Y - X\beta\right\|_2^2 + \varphi(\beta)\right\}, \tag{1}$$

where $X \in \mathbb{R}^{N\times P}$ denotes P feature value from N training cases, $Y \in \mathbb{R}^N$ denotes OAR DVH PCS of cases in the training set, and $\beta \in \mathbb{R}^P$ denotes regression coefficients corresponding to P anatomical features, such as PCS of distance-to-target histogram. Detailed descriptions of feature extraction and dimension reduction for KBP can be found in Ref. (1, 2). The last term, known as the penalty term, balances the bias and variance of the trained model. The goal of KBP is to obtain regression coefficients β based on cases previously planned by experienced planners, and when a new case needs to be planned, the optimal OAR DVH can be calculated simply using the model predicted PCS of $X\beta$. In ridge regression, the penalty term $\varphi(\beta)$ is the square of $\ell 2$-norm of the regression coefficients β; in lasso, the penalty term is the $\ell 1$-norm of β; and in elastic net, the penalty term is simply a linear combination of $\ell 1$-norm and $\ell 2$-norm squared:

$$\varphi(\beta) = \lambda_1 \left\|\beta\right\|_1 + \lambda_2 \left\|\beta\right\|_2^2. \tag{2}$$

The penalty weights λ_1 and λ_1 are selected based on internal cross validation.

Forward selection, a type of stepwise regression, is the last individual model. It finds the most significant features to add based on the data step by step, hence the name. When adding features no longer improves the model by a certain preset *p*-value threshold, the feature selection step terminates. The selected features are fitted to the data with ordinary least square, while the rest of the features are discarded.

The Ensemble Model

Many ensemble models have been proposed over the years in the field of machine learning, such as random forest (17), boosting (18), bagging (19), and stacking (20). The basic idea behind these ensemble models is to develop an array of simple models, often referred to as base learners, and combine these models to form a better (e.g., lower variance, higher accuracy, or both) model for prediction (21). These models essentially seek to combine knowledge learned by different models *via* data resampling and/or adding another layer of optimization.

The primary motivation of our ensemble model is to make KBP more robust and adaptive. In different settings, different regression models perform well, and none of these individual models consistently performs better than other models. For instance, stepwise regression is widely known to be unstable (22), but as shown in Section "Results," it can significantly outperform other more stable models such as ridge regression in certain settings. However, it is not feasible to test out individual models every time a new model is fit. Therefore, we propose an ensemble model, which performs well in all settings.

Model Stacking

In our proposed model, we combine the aforementioned individual models using model stacking method. A previous study demonstrated that even stacking ridge regression alone with different penalty weight λ improved model generalization performance, and stacking models with different characteristics generated further improvement (20). The proposed ensemble approach is shown in Eqs 3–5

$$z_{kn} = \beta_k x_n, k = 1, K, \tag{3}$$

$$\alpha_k^* = \underset{\alpha_k}{\operatorname{argmin}} \sum_{n=1}^{N} \left(y_n - \sum_{k=1}^{K} \alpha_k z_{kn} \right)^2, \text{s.t.} \forall \alpha_k \geq 0, \tag{4}$$

$$Y = \sum_{k=1}^{K} \alpha_k^* \beta_k X. \tag{5}$$

First, individual models β_k, where $k \in [1, K]$ denotes individual model index, are trained separately on the training dataset repetitively with all the training data except for case n. Prediction of the in-training-set but out-of-model case z_{kn} is then generated (Eq. 3). The process is repeated until all the models have covered all cases in the training set. Subsequently, the model weights α_k^* are optimized to minimize internal cross validation error, as shown in Eq. 4. A non-negative constraint is applied to prevent overfitting and increase the model interpretability. This step of optimization is done on the metadata, and the prediction results of each model for each case are used to optimize the model weights. The individual models that perform well in the prediction task tend to get larger weightings. The K individual models β_k are combined and used for prediction of DVH PCS Y (Eq. 5). Note that the sum of optimal model weights α_k^* is not constrained to 1, as one would intuitively expect. This is due to the distinct properties of the individual models in the ensemble. The regression coefficients by stepwise regression are usually too large due to lack of constraint and thus need shrinkage. On the contrary, the other three regression methods tend to under-fit, especially for noisy training data, i.e., data with high variance that cannot be explained by any features in X. In other words, even if we have just one model in the "ensemble," the model weight is still highly unlikely to be 1 (usually smaller than 1 for stepwise and greater than 1 for penalized linear regression methods). In practice, we observe the sum of α_k^* is usually between 0.5 and 1.5.

The ensemble in this study consists of nine models, including stepwise, ridge, lasso, and elastic net with six different λ_2-to-λ_1 ratios. **Figure 1** shows one example of the model weights from the individual models. This model is built using 50 prostate sequential boost cases. Y is the bladder DVH PCS1, and X consists of bladder anatomical features. All features are standardized before training, thus the weights of different features are in the same scale. It is apparent that regression coefficients differ from model to model, even though these are all variants of linear regression models. Note that model 1, stepwise regression, uses the least number of features, and model 2, ridge regression, evidently underfits.

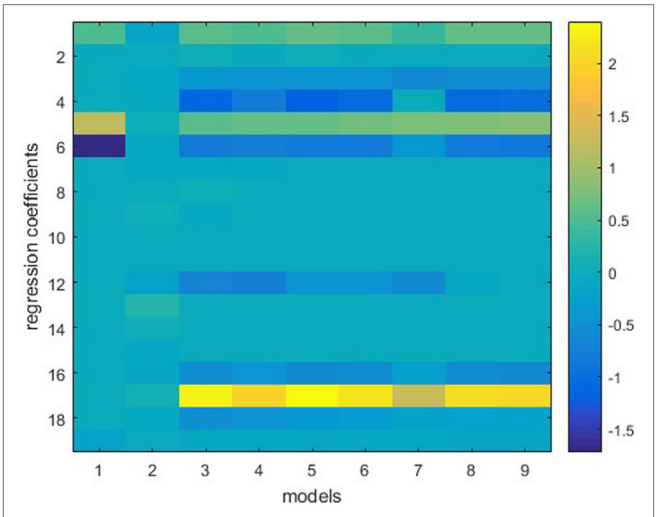

FIGURE 1 | Individual models trained on the same dataset. Vertical lines represent regression coefficients of individual models. Models 1–9 (from left to right) refer to stepwise regression, ridge regression, six elastic net regression models with various parameters, and lasso. The vertical bar on the right indicates color mapping. Note that stepwise regression uses the least (four) features and ridge regression uses all features but assigns small weights to the features.

Model-Based Case Filtering

In previous studies, it has been pointed out that automatic outlier removal requires further investigation (12, 23). We propose to incorporate a model-based automatic outlier removal routine in the ensemble model to ensure model robustness and address the volatile nature of clinical data. We utilize the cross validation metadata native to the proposed ensemble method to identify and remove impactful dosimetric and anatomical outliers. The two scenarios of outliers have different impact on the training of regression models, as we illustrate in this section. Note that by our definition outliers only exist in training sets, all cases in testing sets are predicted. Cases that would be defined as outlier cases if they are in a training set can still be predicted by a trained model, but with less accuracy. These special cases can be identified with the same approach as we identify outlier cases (see Model-Based Case Filtering Method), and case-based reasoning can be used to improve the outcome of treatment planning, but that is out of the scope of this study. We aim to improve prediction accuracy of the KBP framework with a different modeling technique, without significant changes to the overall workflow.

Outliers

Clinical treatment planning varies from case to case, with different sparing and coverage considerations. With the aforementioned KBP framework, we assume a linear model can successfully represent a majority of training cases. For some cases in the database, this assumption does not hold. We refer to these cases in the training dataset as outlier cases. In this section, we shall present our insight on outlier cases and provide an intuitive explanation of effects of outliers on knowledge-based modeling.

Anatomical Outliers and Dosimetric Outliers

The first type of outliers is anatomical outliers. In this study, we define anatomical outliers as cases with anatomical features that are distant from normal cases, and possibly come from a different distribution. In KBP, anatomical outliers refer to cases with uncommon anatomical features relevant to DVH prediction, such as abnormal OAR sizes, unusual OAR volume distributions relative to PTV surface. Generally, anatomical outliers are more likely to deviate from the linear model, as illustrated in **Figure 2**, and when they do, the effect of these cases are generally larger than normal cases due to the quadratic data fidelity term (first term in Eq. 1) of the regression model. Therefore, it is necessary to identify anatomical outlier cases that are detrimental to model building and remove those from the model before training.

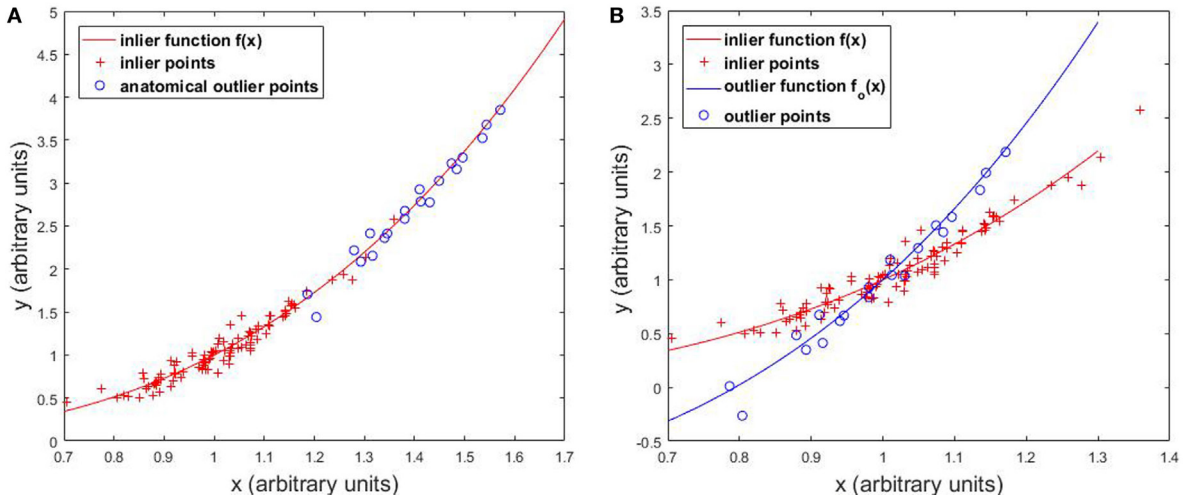

FIGURE 2 | Effects of **(A)** anatomical outliers and **(B)** dosimetric outliers on the regression model.

Other than anatomical outliers, there are cases that are detrimental to model building due to limited OAR sparing efforts and/or capabilities. These are considered to be dosimetric outliers in this work. Dosimetric outliers include, but are not limited to (1) treatment plans with inferior OAR sparing and (2) wrongly labeled data, such as 3D plans mixed in IMRT plans.

Outliers' Effect on Regression Models

In this section, we illustrate the effect of outliers on the overall regression model with one-dimensional simulated data. **Figure 2A** shows that anatomical outliers follow the same underlying X-to-Y mapping. However, the true underlying relation may not be well approximated by linear regression outside the normal X range. Attempting to fit linear regression with anatomical outliers mixed in the training set will potentially deteriorate the model. Therefore, the actual effect of anatomical outlier in different feature directions in the context of KBP needs careful assessment. **Figure 2B** illustrates the effect of dosimetric outliers. Dosimetric outliers in the training set are expected to increase model variance and deviate the model.

Note that this numerical demonstration isolates the effect of outliers on regression on a single feature, and it simplifies the influence of outliers on the overall modeling process. In our clinical knowledge-based modeling, we extract nine features from each case to construct the feature vector X. However, not every feature contributes to the final model equally. In stepwise regression, relevant features are picked based on correlation with the outcomes variable (i.e., DVH PCS). In penalized regression methods, features are implicitly selected with less relevant features given very small regression coefficients as a result of the penalty term. The feature selection step, while not considered here, is also affected by outliers. When anatomical outliers are involved in the training process, the features selected are potentially different from the set of features selected, if the model is trained without outliers.

Prediction Performance Measure

Weighted root mean squared error (wRMSE) is defined to evaluate model prediction accuracy:

$$\text{wRMSE} = \sum_{i=1}^{N} w_i' \left(\text{DVH}_i - \widehat{\text{DVH}_i}\right)^2. \tag{6}$$

Weighted root mean squared error measures the overall deviation of predicted DVHs from ground truth DVHs, which are clinically planned. Weightings are introduced to emphasize higher dose regions of DVHs, which are generally considered to be of more clinical significance in OAR dose predictions. Here $w_i' = Nw_i / \sum_{j=1}^{N} w_j$ denotes the normalized weighting factor for bin i of DVH curves. For evaluation of dose to bladder and rectum, we use the linear relative weighting w_j of 50–100 linearly increases from 0 Gy to prescription dose. For evaluation of dose to parotids in head and neck cases, w_i is set to Gaussian centered at median dose, with SD of 2 Gy. If w_i is set to a constant number, then wRMSE reduces to standard RMSE.

Model-Based Case Filtering Method

To further improve the robustness of the ensemble model, cases with the highest s% median (of all individual models) internal cross validation wRMSE error are dropped from the training set. The percentage threshold s is selected to balance the tradeoff between model robustness and accuracy. Empirically, we find that 10% is generally a good choice, even though the number of actual outlier cases is unknown and may differ from 10% of the total case number. All the experiments in the following section are conducted with the pre-determined 10% threshold. The workflow of the ensemble model with model-based case filtering is shown in **Figure 3**. Note that the whole process is done automatically without manual intervention.

Experimental Design

This retrospective study uses anonymized clinical plan data and has received permission from Duke University Medical Center's

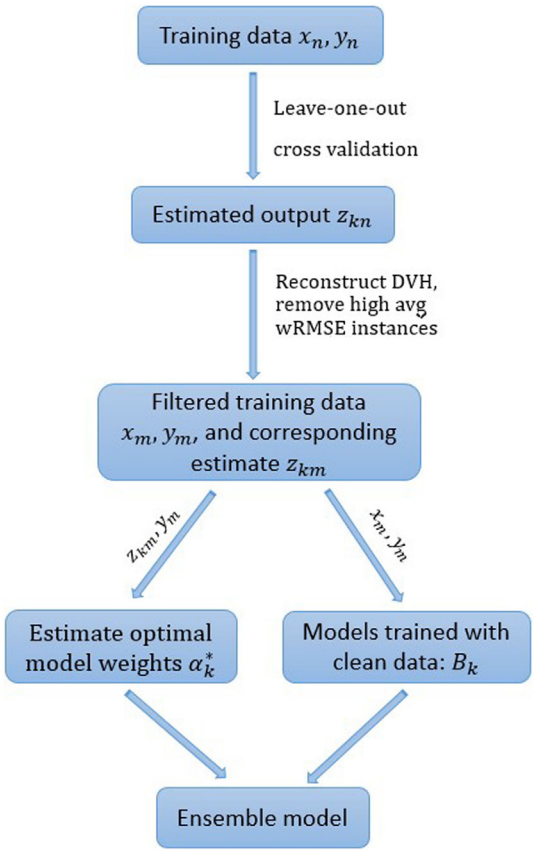

FIGURE 3 | The proposed ensemble learning workflow.

institutional IRB. All clinical plans were planned using Varian Eclipse™ Treatment Planning System (Varian Medical Systems, Inc., Palo Alto, CA, USA). All experiments were performed on a PC with Intel Xeon E5-2623 CPU and 32 GB of RAM running Windows 10 Enterprise 64-bit operating system.

In order to quantitatively evaluate the robustness of these regression methods in various challenging clinical environment, we test the aforementioned models with limited training set size, training sets contaminated with anatomical outliers, and training sets contaminated with dosimetric outliers. In our outlier robustness tests, we purposefully mix pre-defined outlier cases into the training set and validate the final model with normal cases. The reason for adding outlier cases is to add controlled variation to the dataset and evaluate the robustness of the proposed model. Details regarding types of data used in the experiments are summarized in **Table 1**.

Robustness to Limited Training Set Size

In clinical practice, planners do not necessarily have many cases for every treatment site. This is particularly true when a new treatment technique, such as simultaneous intensity boost, is recently utilized in the clinic and the existing model built for existing treatment techniques may not predict the achievable DVH accurately due to the OAR sparing capability difference. Sometimes models need to be built when only a small number of cases (~20) are available. It is critical that the regression model is capable of resisting overfitting the random variation of training cases. In this experiment, 166 prostate PTV cases are retrospectively retrieved from the clinical database. Twenty prostate cases are used as the training set, and the remaining 146 cases are used as validation set to quantitatively evaluate the prediction accuracy of each model.

TABLE 1 | Summary of data used in the experiments.

Experiments	Training data	Validation data
Limited training set size	20 prostate intensity-modulated radiation therapy (IMRT) cases	146 prostate IMRT cases
Anatomical outliers	10 prostate cases treated with lymph nodes and 40 prostate cases treated without lymph node	111 prostate cases treated without lymph node
Dosimetric outliers (inferior plans)	40 prostate IMRT cases and 10 prostate conformal arc plans	110 prostate IMRT plans
Dosimetric outliers (mis-classified sparing decisions)	80 bilateral parotid-sparing head and neck plans and 10 single-side sparing plans	148 bilateral parotid-sparing head and neck plans

Robustness to Anatomical Outliers

In clinical databases, not every previously treated case is helpful for predicting future cases even when the treatment plans are of high quality. If the anatomical features are very different from the majority of all cases than the linear assumption may not hold, as demonstrated in **Figure 2**, and the anatomical features are potentially detrimental to the model. To simulate the effect of anatomical outliers on the plans, we train a model with 10 prostate cases treated with lymph nodes and 40 prostate cases treated without lymph node. The trained models are subsequently validated with 111 cases that do not involve lymph nodes.

Robustness to Dosimetric Outliers

Dosimetric outliers do not follow the same conditional distribution as normal cases and are expected to be easier to be identified with cross validation. Increase of dosimetric outliers in training data tends to shift the overall model toward inferior plan DVHs and gradually make the plan less optimal (23). In this section, we evaluate the robustness of individual models and the ensemble model with training set contaminated by two types of dosimetric outlier plans: (i) inferior dose sparing and (ii) mis-labeled sparing decisions.

For KBP, it is crucial to get reliable predictions even in the presence of sub-optimal plans. Here, we simulate the sub-optimal plans with dynamic conformal arc plans. Compared with IMRT plans, conformal arc plans have evidently inferior OAR sparing capability. Our training data consists of 40 prostate IMRT cases and 10 prostate conformal arc plans, and the validation set includes 110 prostate IMRT plans. The experiment is designed to test the model robustness in the extreme settings to evaluate the model robustness in challenging situations.

In clinical practice, it is not always feasible to spare both parotids due to geometric factors. A previous study has shown that parotid-sparing decisions affect KBP predictions, and separate models should be built for single-side parotid sparing and

bilateral parotid sparing to get better prediction accuracy (24). We retrieve 228 bilateral parotid-sparing head and neck cases and 10 single-side parotid-sparing cases from our institutional clinical database. The sparing decisions are first obtained from clinical prescription documentations and subsequently checked in dose statistics to correct for decision changes. We randomly select 80 bilateral cases as the training set and then add 10 single-side sparing cases as mis-classified cases. The remaining 148 bilateral cases are used as the validation set.

RESULTS

Robustness to Limited Training Set Size

The ensemble method outperforms all individual methods significantly, as shown in **Figure 4**. Note that ridge regression performs particularly poorly in bladder prediction, indicating that there is some intrinsic sparsity in the feature space, and ridge regression, which does not utilize that sparsity, underfits significantly due to over-shrinking of regression coefficients. Stepwise performs poorly in rectum predictions, due to overfitting.

Robustness to Anatomical Outliers

Figure 5 shows prediction errors, measured by wRMSE, of individual models and the ensemble model. For bladder predictions, the ensemble model outperforms all individual models, while stepwise, lasso, and elastic net perform similarly. In the case of rectum predictions, the ensemble method again outperforms ridge, lasso, and elastic net, and performs similarly well as stepwise. Ridge regression fails to predict accurately for either task.

Robustness to Dosimetric Outliers

Inferior Plans

Figure 6 shows, for both bladder and rectum prediction, lasso, elastic net, and the proposed ensemble regression method predict

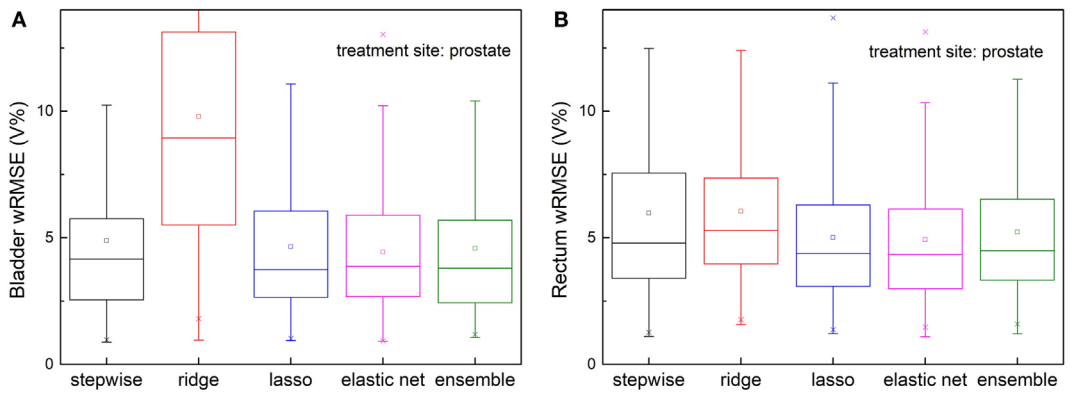

FIGURE 4 | Prediction errors [weighted root mean squared error (wRMSE)] of individual regression models and the proposed ensemble model for **(A)** bladder **(B)** rectum. The training set and validation set for all the models tested are identical. Twenty prostate 1PTV cases are included in training set, and the validation set includes 146 cases. For bladder prediction, the proposed ensemble method predicts significantly better than stepwise ($p < 0.001$), ridge ($p < 0.001$), lasso ($p < 0.001$), and elastic net ($p < 0.001$); for rectum prediction, the proposed ensemble method predicts significantly better than ridge ($p < 0.001$), lasso ($p < 0.001$), elastic net ($p < 0.001$), and stepwise ($p < 0.001$).

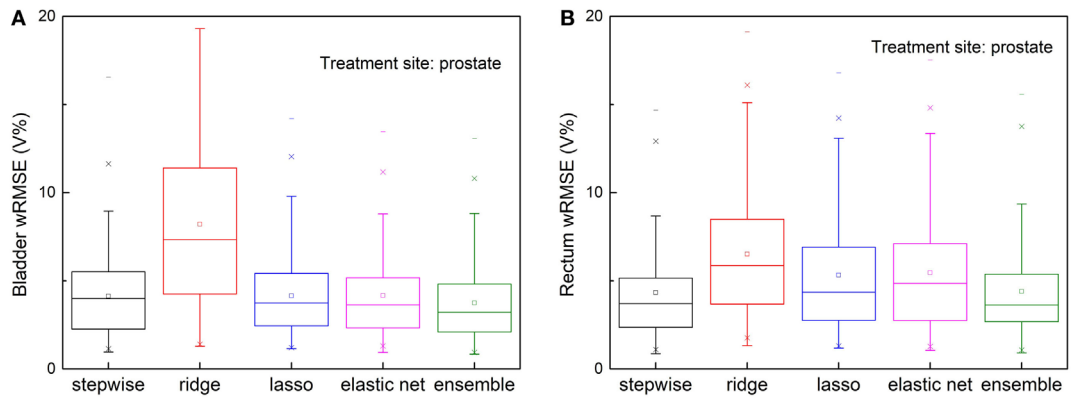

FIGURE 5 | Prediction errors [Weighted root mean squared error (wRMSE)] of individual regression models and the proposed ensemble model for **(A)** bladder **(B)** rectum, in the presence of simulated anatomical outliers (see text). Forty prostate with seminal vesicle cases and 10 prostate cases with lymph node cases are used as the training set; 111 cases prostate with seminal vesicle cases are used as the validation set. For bladder prediction, the proposed ensemble method predicts significantly better than stepwise ($p = 0.013$), ridge ($p < 0.001$), lasso ($p = 0.002$), and elastic net ($p < 0.001$); for rectum prediction, the proposed ensemble method predicts significantly better than ridge ($p < 0.001$), lasso ($p < 0.001$), elastic net ($p < 0.001$), and performs similarly well as stepwise ($p = 0.210$).

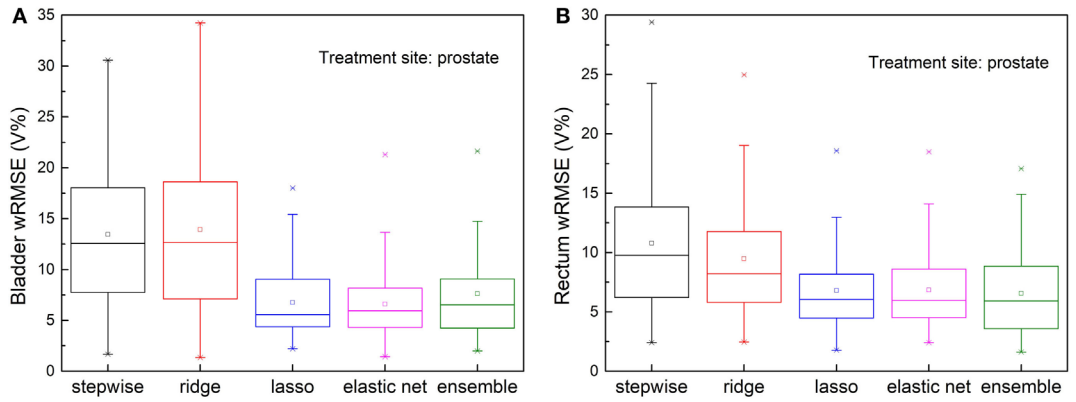

FIGURE 6 | Prediction errors [Weighted root mean squared error (wRMSE)] of individual regression models and the proposed ensemble model for **(A)** bladder **(B)** rectum, in the presence of simulated dosimetric outliers (see text). For bladder prediction, the proposed ensemble method predicts significantly better than stepwise ($p < 0.001$), ridge ($p < 0.001$), and performs similarly well as lasso ($p = 0.753$) and elastic net ($p = 0.841$). For rectum prediction, the proposed ensemble method predicts significantly better than stepwise ($p < 0.001$) and ridge ($p < 0.001$), and performs similarly well as lasso ($p = 0.365$) and elastic net ($p = 0.373$).

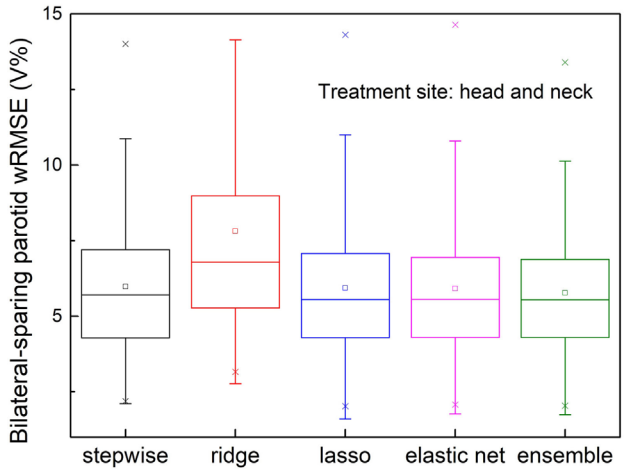

FIGURE 7 | Prediction errors [Weighted root mean squared error (wRMSE)] of individual regression models and the proposed ensemble models for parotid dose volume histogram (DVH) prediction of bilateral parotid-sparing cases. Training set includes 80 bilateral sparing cases and 10 single-side sparing cases. Validation set includes 148 bilateral sparing cases. The proposed ensemble model yields significantly reduced prediction error than stepwise ($p = 0.026$) and ridge ($p < 0.001$), but does not outperform elastic net ($p = 0.090$) or lasso ($p = 0.115$).

equally well, while stepwise and ridge are no longer usable due to significant amount of error.

Mis-Classified Sparing Decisions

The validation set prediction errors of each model are shown in **Figure 7**. The proposed ensemble model significantly reduces prediction error, compared with stepwise ($p = 0.026$) and ridge ($p < 0.001$), and performs equally well as elastic net ($p = 0.091$) and lasso ($p = 0.115$).

DISCUSSION

In summary, we propose an ensemble regression model to address two problems that we are facing in KBP. First, different individual regression models perform well in different settings, such as different number of relevant features, number of cases, and existence of outliers. It would be very labor intensive to manually select the optimal model every time a model is fitted. Second, to ensure the most accurate model training, data-preprocessing, including anatomical and dosimetric outlier removal, is also necessary for individual models, and it can be subjective to decide which subset of cases should be removed from the training set if done manually. The proposed ensemble model utilizes multiple individual models on the same set of data and uses constrained linear optimization on the metadata to obtain the optimal weight for each individual model. In addition, the model automatically filters out cases in the training set that are not predicative of future cases based on metadata.

We observe that the ensemble method consistently predicts better than or similar to the best performing individual model in every challenging situation. With improved robustness, the proposed regression method potentially enables end users to build site-specific, physician-specific, or even planner specific models, without manually screening the training cases. This eventually will allow each practice to build models that accurately reflect their own optimal OAR sparing preference and capability, thereby eliminating the need for a universal model.

Figure 8 shows an example of improved prediction accuracy of the proposed method, compared with other individual models. In this case, stepwise and ridge perform poorly while lasso and elastic net perform reasonably well, and the ensemble model outperforms all individual models. Note that in different situations, different models perform well, and the proposed model performs most consistently. Improved DVH prediction accuracy usually results in better plan optimization guidance (i.e., optimization

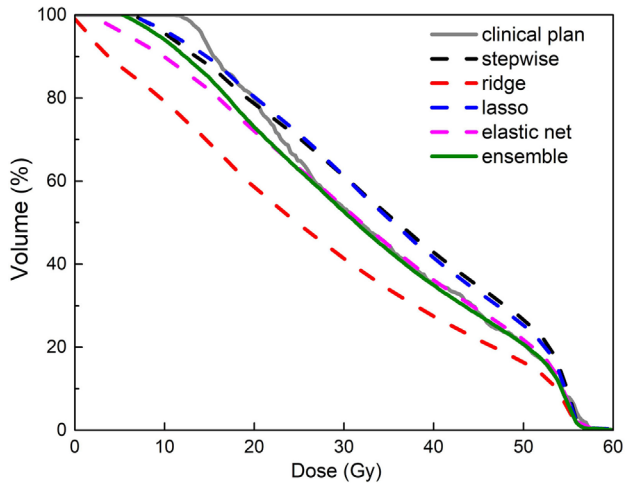

FIGURE 8 | An example of improved accuracy of the ensemble model (green solid line) in predicting a bladder dose volume histogram (DVH) for a prostate plan over individual models (dashed lines in other colors). All models were trained with data that include dosimetric outliers (see Robustness to Dosimetric Outliers). The clinical plan DVH (gray solid line) is the "ground truth." Note that the green line follows the gray line most closely.

constraint generation), since it provides the treatment planning system correct information of the best achievable OAR sparing without compromising PTV coverage.

Building models for different treatment sites may face different challenges. For example, the number of cases required to train a model may be different. The more complex head and neck cases require more training cases to well represent the case population, while prostate cases have fewer OARs and are generally easier to train. Second, different treatment strategies are often used to treat different sites. For example, some sites require multiple PTVs while other sites require hard constraints. Last but not least, the amount of intrinsic variance in head and neck cases are more than that of prostate cases due to potential trade-off considerations. As a result, dataset characteristics vary from treatment site to treatment site and individual model performances vary correspondingly. The ensemble model ensures the best performing model gets the highest weighting. All in all, each treatment site should be treated differently in KBP to get the best possible prediction accuracy, and the ensemble model helps to reduce the amount of effort required in terms of model selection. Ideally, the ensemble method should be trained for each treatment site, since data characteristics change from dataset to dataset. However, if there are two datasets from two treatment sites with very similar characteristics, such as DVH variability, number of cases, then it is possible to re-use the model weight α_k^* directly.

The main limitation of the proposed approach is the training time. Two major components of knowledge-based modeling are feature extraction and model training. The feature extraction part of the proposed model takes on average 5 s for each case, and feature extraction is done only once. Model training takes less than 10 s for each individual model. In the proposed model, individual model training is repeated by the number of component models times the number of in-model cross validation. As a result, in our hardware setup, it takes less than 10 min to run a single regression model, and it takes 30 min to run a 20-fold cross-validated ensemble model. The prediction procedure is very simple and takes less than 1 s to calculate Therefore, once a model is calculated, it can be easily stored and applied to DVH predictions.

Possible future research topics include the optimal selection of models as well as the optimal number of models in the ensemble. In this study, we limit the number of models included in the training set to avoid overfitting. While too many models in the ensemble warrant overfitting the data, the current number of models (9) is very conservative. With the regulation of the non-negative constraint, the proposed approach could potentially see further performance improvements if more models are included in the ensemble. We expect the optimal number of models in the ensemble to be dependent of the size of the dataset. In addition, the proposed methodology can be easily expanded to more complicated non-linear models. We use linear models in the ensemble due to the limitations of training dataset size. As more cases become available, more complicated models become viable.

AUTHOR CONTRIBUTIONS

JZ proposed the model, conducted experiments, and wrote the first draft of the manuscript. QW oversaw the workflow of the study and contributed in the clinical aspect of the study. TX extracted and pre-processed data for the experiments in the paper. YS provided suggestions regarding the study design. F-FY provided critics in the experimental design. YG contributed advice in the statistical methods and revised the manuscript.

FUNDING

This work is partially supported by NIH under grant #R01CA-201212 and a master research grant by Varian Medical Systems.

REFERENCES

1. Zhu X, Ge Y, Li T, Thongphiew D, Yin FF, Wu QJ. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Med Phys* (2011) 38(2):719–26. doi:10.1118/1.3539749
2. Yuan L, Ge Y, Lee WR, Yin FF, Kirkpatrick JP, Wu QJ. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. *Med Phys* (2012) 39(11):6868–78. doi:10.1118/1.4757927
3. Appenzoller LM, Michalski JM, Thorstad WL, Mutic S, Moore KL. Predicting dose-volume histograms for organs-at-risk in IMRT planning. *Med Phys* (2012) 39(12):7446–61. doi:10.1118/1.4761864
4. Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys* (2011) 81(2):545–51. doi:10.1016/j.ijrobp.2010.11.030
5. Wu B, Ricchetti F, Sanguineti G, Kazhdan M, Simari P, Jacques R, et al. Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys* (2011) 79(4):1241–7. doi:10.1016/j.ijrobp.2010.05.026
6. Hussein M, South CP, Barry MA, Adams EJ, Jordan TJ, Stewart AJ, et al. Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy. *Radiother Oncol* (2016) 120(3):473–9. doi:10.1016/j.radonc.2016.06.022

7. Wu H, Jiang F, Yue H, Zhang H, Wang K, Zhang Y. Applying a RapidPlan model trained on a technique and orientation to another: a feasibility and dosimetric evaluation. *Radiat Oncol* (2016) 11(1):108. doi:10.1186/s13014-016-0684-9
8. Fogliata A, Belosi F, Clivio A, Navarria P, Nicolini G, Scorsetti M, et al. On the pre-clinical validation of a commercial model-based optimisation engine: application to volumetric modulated arc therapy for patients with lung or prostate cancer. *Radiother Oncol* (2014) 113(3):385–91. doi:10.1016/j.radonc.2014.11.009
9. Tol JP, Dahele M, Delaney AR, Slotman BJ, Verbakel WF. Can knowledge-based DVH predictions be used for automated, individualized quality assurance of radiotherapy treatment plans? *Radiat Oncol* (2015) 10:234. doi:10.1186/s13014-015-0542-1
10. Berry SL, Ma R, Boczkowski A, Jackson A, Zhang P, Hunt M. Evaluating inter-campus plan consistency using a knowledge based planning model. *Radiother Oncol* (2016) 120(2):349–55. doi:10.1016/j.radonc.2016.06.010
11. Chang AT, Hung AW, Cheung FW, Lee MC, Chan OS, Philips H, et al. Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy. *Int J Radiat Oncol Biol Phys* (2016) 95(3):981–90. doi:10.1016/j.ijrobp.2016.02.017
12. Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WFAR. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys* (2015) 91(3):612–20. doi:10.1016/j.ijrobp.2014.11.014
13. Tikhonov AN. On the stability of inverse problems. *Cr Acad Sci Urss* (1943) 39:176–9.
14. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* (2000) 42(1):80–6. doi:10.2307/1271436
15. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B* (1996) 58(1):267–88.
16. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* (2005) 67:301–20. doi:10.1111/j.1467-9868.2005.00503.x
17. Breiman L. Random forests. *Mach Learn* (2001) 45(1):5–32. doi:10.1023/a:1010933404324
18. Schapire RE. The strength of weak learnability. *Mach Learn* (1990) 5(2):197–227. doi:10.1023/a:1022648800760
19. Breiman L. Bagging predictors. *Mach Learn* (1996) 24(2):123–40. doi:10.1007/bf00058655
20. Wolpert DH. Stacked generalization. *Neural Netw* (1992) 5(2):241–59. doi:10.1016/S0893-6080(05)80023-1
21. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer-Verlag (2009).
22. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat* (1996) 24(6):2350–83. doi:10.1214/aos/1032181158
23. Delaney AR, Tol JP, Dahele M, Cuijpers J, Slotman BJ, Verbakel WFAR. Effect of dosimetric outliers on the performance of a commercial knowledge-based planning solution. *Int J Radiat Oncol Biol Phys* (2016) 94(3):469–77. doi:10.1016/j.ijrobp.2015.11.011
24. Yuan L, Wu QJ, Yin F-F, Jiang Y, Yoo D, Ge Y. Incorporating single-side sparing in models for predicting parotid dose sparing in head and neck IMRT. *Med Phys* (2014) 41(2):021728. doi:10.1118/1.4862075

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 16 April 2018
doi: 10.3389/fonc.2018.00108

Lung Nodule Detection *via* Deep Reinforcement Learning

*Issa Ali[1,2], Gregory R. Hart[1], Gowthaman Gunabushanam[3], Ying Liang[1], Wazir Muhammad[1], Bradley Nartowt[1], Michael Kane[4], Xiaomei Ma[2] and Jun Deng[1]**

[1] Department of Therapeutic Radiology, School of Medicine, Yale University, New Haven, CT, United States, [2] Department of Chronic Disease Epidemiology, School of Public Health, Yale University, New Haven, CT, United States, [3] Department of Radiology and Biomedical Imaging, School of Medicine, Yale University, New Haven, CT, United States, [4] Department of Biostatistics, School of Public Health, Yale University, New Haven, CT, United States

Lung cancer is the most common cause of cancer-related death globally. As a preventive measure, the United States Preventive Services Task Force (USPSTF) recommends annual screening of high risk individuals with low-dose computed tomography (CT). The resulting volume of CT scans from millions of people will pose a significant challenge for radiologists to interpret. To fill this gap, computer-aided detection (CAD) algorithms may prove to be the most promising solution. A crucial first step in the analysis of lung cancer screening results using CAD is the detection of pulmonary nodules, which may represent early-stage lung cancer. The objective of this work is to develop and validate a reinforcement learning model based on deep artificial neural networks for early detection of lung nodules in thoracic CT images. Inspired by the AlphaGo system, our deep learning algorithm takes a raw CT image as input and views it as a collection of states, and output a classification of whether a nodule is present or not. The dataset used to train our model is the LIDC/IDRI database hosted by the lung nodule analysis (LUNA) challenge. In total, there are 888 CT scans with annotations based on agreement from at least three out of four radiologists. As a result, there are 590 individuals having one or more nodules, and 298 having none. Our training results yielded an overall accuracy of 99.1% [sensitivity 99.2%, specificity 99.1%, positive predictive value (PPV) 99.1%, negative predictive value (NPV) 99.2%]. In our test, the results yielded an overall accuracy of 64.4% (sensitivity 58.9%, specificity 55.3%, PPV 54.2%, and NPV 60.0%). These early results show promise in solving the major issue of false positives in CT screening of lung nodules, and may help to save unnecessary follow-up tests and expenditures.

Keywords: lung cancer, computed tomography, lung nodules, computer-aided detection, reinforcement learning

OPEN ACCESS

Edited by:
Radka Stoyanova, University of Miami, United States

Reviewed by:
Patrik Brodin, Albert Einstein College of Medicine, United States
Bilgin Kadri Aribas, Bülent Ecevit University School of Medicine, Turkey

****Correspondence:***
Jun Deng
jun.deng@yale.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *29 January 2018*
Accepted: *28 March 2018*
Published: *16 April 2018*

Citation:
Ali I, Hart GR, Gunabushanam G, Liang Y, Muhammad W, Nartowt B, Kane M, Ma X and Deng J (2018) Lung Nodule Detection via Deep Reinforcement Learning. Front. Oncol. 8:108. doi: 10.3389/fonc.2018.00108

INTRODUCTION

Computed tomography (CT) is an imaging procedure that utilizes X-rays to create detailed images of internal body structures. Presently, CT imaging is the most preferred method to screen the early-stage lung cancers in at-risk groups (1). Globally, lung cancer is the leading cause of cancer-related death (2). In the United States, lung cancer strikes 225,000 people every year and accounts for $12 billion in healthcare costs (3). Early detection is critical to give patients the best chance of survival and recovery.

Screening high risk individuals with low-dose CT scans has been shown to reduce mortality (4). However, there is significant inter-observer variability in interpreting screenings as well as a large number of false positives which increase the cost and reduce the effectiveness of screening programs. Given the high incidence of lung cancer, optimizing screening by reducing false positives and false negatives has significant public health impact by limiting unnecessary biopsies, radiation exposure, and other secondary costs of screening (5).

Several studies have shown that imaging can predict lung nodule presence to a high degree (6). Clinically, detecting lung nodules is a vital first step in the analysis of lung cancer screening results—the nodules may or may not represent early-stage lung cancer. Numerous computer-aided detection (CAD) methods have been proposed for this task. The majority, if not all, utilize classical machine learning approaches such as supervised/unsupervised methods (7). The goal of this work is to adopt for the first time a reinforcement learning (RL) algorithm for lung nodule detection. Developed by Google DeepMind, RL is a cutting-edge machine learning approach which has improved upon numerous CAD systems and helped to beat the best human players in the game of Go, one of the most complex games humans ever invented (8). Here, we apply RL to the lung nodule analysis (LUNA) dataset and analyze the performance of the RL model in detecting lung nodules from thoracic CT images.

MATERIALS AND METHODS

Lung Nodule Data

For the training of our algorithm, we utilize the LUNA dataset, which curates CT images from publicly available LIDC/IDRI database. In total, there are 888 CT scans included. The database also contains annotations collected in two phases with four experienced radiologists. Each radiologist marked lesions they identified as non-nodule (<3 mm) and nodule (≥3 mm) and the annotation process has been described previously (9). The reference standard consists of all nodule ≥3 mm accepted by at least three out of four radiologists. Annotations that are not included in the reference standard (non-nodules, nodules <3 mm, and nodules annotated by only one or two radiologists) are referred to as irrelevant findings (9). A key benefit of this dataset is the inclusion of voxel coordinates in the annotation of nodules, which proves immensely useful when using a RL approach, described in the next section. **Figure 1** illustrates examples of nodule and non-nodules from a single CT scan.

Data Normalization

To balance the intensity values and reduce the effects of artifacts and different contrast values between CT images, we normalize our dataset. The *Z* score for each image is calculated by subtracting the mean pixel intensity of all our CT images, μ, from each image, *X*, and dividing it by σ, the SD of all images' pixel intensities. This step is helpful when inputting information into a neural network because it fine-tunes the input information fed into a convolution algorithm (10).

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Reinforcement Learning

Reinforcement learning is the science of mapping situations to actions (11). It is a type of machine learning that bridges the well-established classical approaches of supervised and unsupervised learning, where target values are known and unknown, respectively. RL differs in that it seeks to model data without any labels, but rather with incremental feedback. Its recent popularity stems from its ability to develop novel solution schemas, even outperforming humans in certain domains, because it learns to solve a task by itself (12). Essentially, it is a way of programming agents by either a reward or a punishment without the need to specify how a task is to be achieved. A simple RL model is shown in **Figure 2** illustrating how an agent's actions in a given environment affect its resulting reward and state. In its infancy, RL was inspired by behavioral psychology, where agents (i.e., rodent) learned tasks by being given a reward for a correct action taken in a given state. This mechanism ultimately creates a feedback loop. Whether the agent, in our case a neural network model, navigates a maze, plays a game of ping pong, or detects lung nodules, the approach is the same.

A basic reinforcement algorithm is modeled after a Markov decision process. For a set number of states, there are a given number of possible actions, and a range of possible rewards (13). To help optimize an agent's actions a *Q*-learning algorithm is used (14).

$$Q(s_t, a_t) = Q(s_t, a_t) + l^* \left[r_{t+1} + \max Q(s_{t+1}, a_t) - Q(s_t, a_t) \right] \quad (2)$$

How a model knows the potential rewards from taking a certain action comes from experience play. That is, it stores numerous combinations of state to state transitions ($s \rightarrow s^{+1}$), with the corresponding action, *a*, taken by the model and the resultant reward, *r*: denoted as (s, a, r, s^{+1}). For instance, in a game environment, the best action to take would be the action that leads to the greatest future rewards (i.e., winning the game), even though the most immediate action may not be rewarding in the short term. As shown in Eq. 2, the expected future rewards are approximated by multiplying the discount rate, λ, by the value of the action that would return the largest future reward based on all possible actions, $\max Q(s_{t+1}, a_t)$. For a given action, what is learned is the reward for that action, r_{t+1}, plus the largest future reward expected less current action value, $Q(s_t, a_t)$. This is learned at a rate, *r*, the extent to which the algorithm overrides old information, and it is valued between 0 and 1. To learn which series of actions result in the greatest number of future rewards, RL algorithms depend on both greedy and exploratory search. The two methods allow a model to explore all possible ways to accomplish a task, and select the most efficient rewarding (12).

Using the RL approach to tackle the lung nodule task requires one main adaptation, which is how we define a state. In a typical RL task, a state would refer to a snapshot of everything

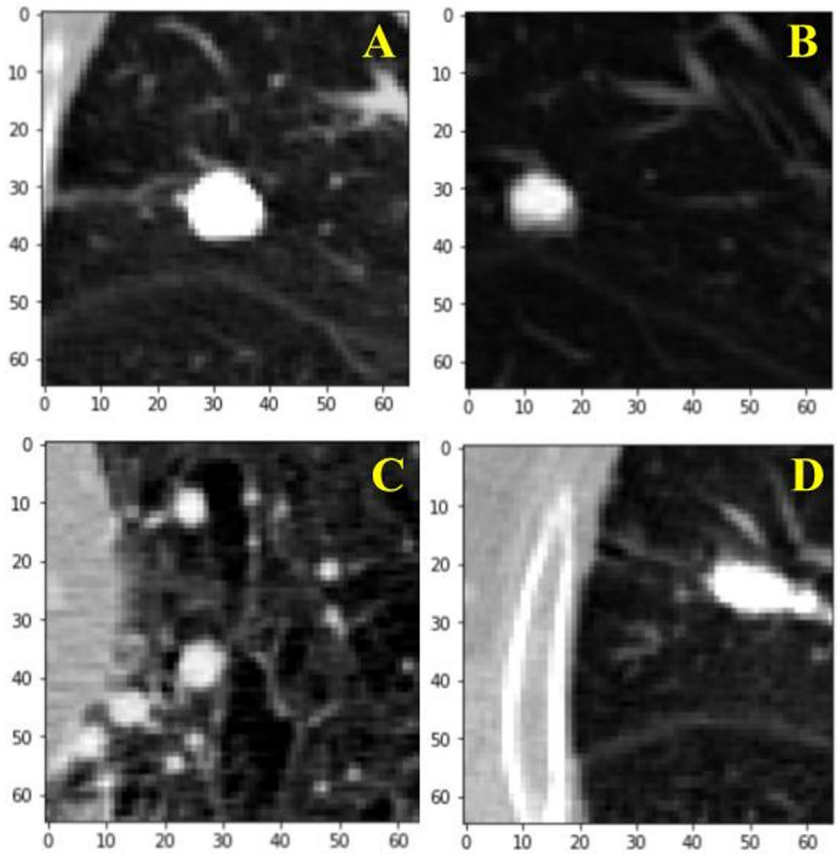

FIGURE 1 | Visual illustration of a sample nodule and non-nodule structure in the lung nodule analysis dataset. Frame **(A)** is a nodule. Frames **(B–D)** are non-nodules.

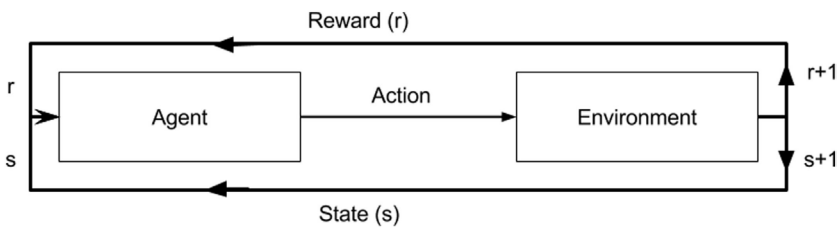

FIGURE 2 | A diagram of a reinforcement model. An agent in a given state (s) and reward (r) completes an action in environment. This results in change of environment and either an increase/decrease in reward as a result of that action.

in an environment at a certain time. However, with lung CT images, which are a collection of axial lung scans, we define a *state* as every 10 stacked axial images. Hence, our environment is very deterministic. That is, any *action* taken in a lung CT image state would lead to the succeeding 10 scans, from top to bottom. Whereas in a conventional task, such as playing a game, depending on the action there is more than one succeeding state possible. This key difference adapts our reward function to act solely as a reward function and evaluate a state on whether it immediately has a reward or not, instead of incorporating a value function which factors the total reward our agent can expect from a given state in the distant future. This makes logical sense given that there is only one possible distant future in our radiographic image environment, whereas in a game environment there is more than one possible distant future. As such, rewards are 1 and 0, depending on whether a classification is correct or incorrect, respectively, for the immediate state at hand only. Thus, the memory replay used to train our model, excludes the succeeding state, and only captures current state, action, and reward.

Convolutional Neural Networks (CNNs)

Learning to control agents directly from high dimensional sensory inputs (i.e., vision and speech) is a significant challenge in RL (11). A key component of our RL model is a CNN. It helps our model make sense of the very high dimensional CT images that we insert into our model. A standard slice has a width and length of 512 × 512. With our input of 10 slices for every state, this amounts to approximately 2,621,440 pixels. A CNN is able to contend with this because it creates a hierarchical representation of high dimensional data such as an image (10).

Unlike a regular neural network, the layers of a CNN have neurons arranged in three dimensions (width, height, and depth) and respond to a receptive field, a small region of the input image, as opposed to a fully connected layer which responds to all the neurons. For a given neuron, it learns to detect features from a local region, which facilitates the capturing of local structures while preserving the topology of the image. The final output layer reduces the image into a vector of class scores. A CNN deep learning system is composed of five layers: an input layer, a convolutional layer, an activation layer, a pooling layer, and a fully connected layer. With most CNN architectures having more than one of each layer, they are thus referred to as "deep" learning (10). The function of each layer is described further below.

Input Layer

This layer holds the raw pixels values of the input image (colored blue in **Figure 3**).

Convolutional Layer

This layer visualized by the red boxes in **Figure 3** is composed of several feature maps along the depth dimension, each corresponding to a different convolution filter. All neurons with the same spatial dimension are connected to the same receptive field of the input image. This facilitates capturing a wide variety of imaging features. The depth of the layer, meaning the number of convolution filters, represents the number of features that can be extracted from each input receptive field. Each neuron in a feature map shares exactly the same weights, which define the convolution filter. This allows reducing the number of weights, and thus increasing the generalizability of the architecture (10).

Activation Layer

Often seen as one with the convolutional layer, as in **Figure 3**, the activation layer applies a threshold function to the output of each neuron in the previous layer. In our network, we use a rectified linear unit (RELU) activation, where $\mathrm{RELU}(x) = \max(0, x)$, meaning it fires the real value of the output and thresholds at zero. It simply replaces the negative values with "0."

Pooling Layer

Typically placed after an activation layer, this layer down-samples along spatial dimensions. Shown by the purple box in **Figure 3**, it selects the invariant imaging features by reducing the spatial dimension of the convolution layer. The most commonly used is max pooling, which selects the maximum value of four of its inputs as the output, thus preserving the most prominent filter responses.

Fully Connected Layer

Shown as green in **Figure 3**, this layer connects all neurons in the previous layer with a weight for each connection. As the output layer, each output nodes represents the "score" for each class.

To facilitate the learning of complex relationships, multiple convolutional-pooling layers are combined to form a deep architecture of nonlinear transformations, helping to create a hierarchical representation of an image. This allows learning complex features with predictive power for image classification tasks (10). As illustrated in **Figure 3**, we use 3D CNN given that nodules are spherical in shape, and can best be captured with 3D convolutions.

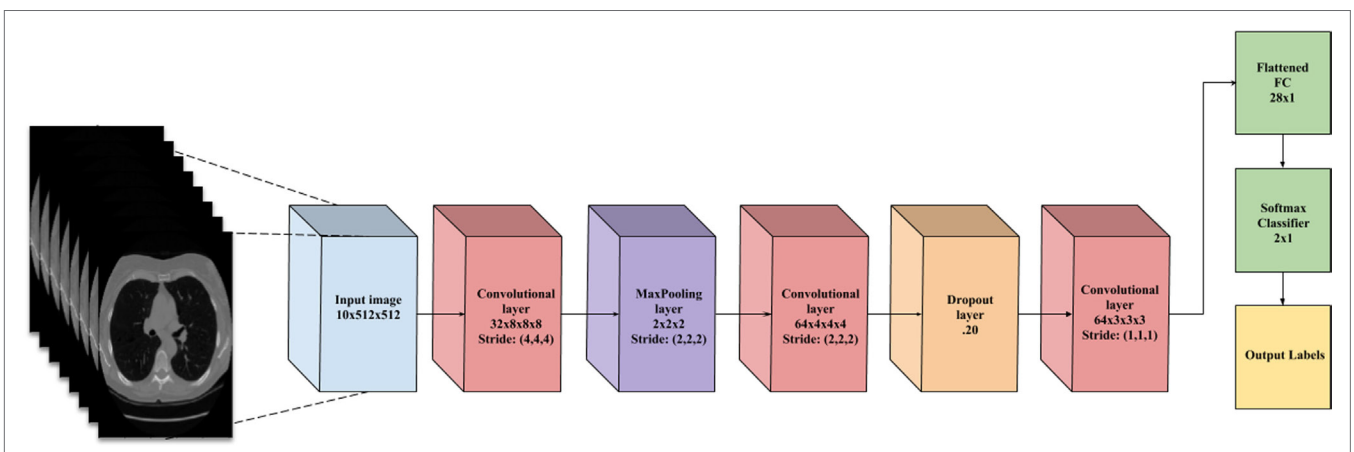

FIGURE 3 | A flowchart of the convolutional neural network architecture. Blue box is the input image. Red boxes are convolutional layers with rectified linear unit activation. Purple box is the max pooling layer. Cyan box is the dropout layer. Green box is the fully connected layer and softmax binary classifier. Yellow is the output of the network.

Data Augmentation

Overfitting is a result of network parameters greatly outnumbering the number of features in the input images. Given the network size and the number of features available from the CT images, our model tended to overfit, hence the need to increase the number of CT images. To counter this overfitting, we used standard deep neural network methods, such as artificially augmenting the dataset using label-preserving transformations (15). The data augmentation consists of applying various image translations, such as rotations, horizontal and vertical flipping, and inversions. We apply a random combination of these transformations on each image, thus creating nominally "new" images. This multiplies the dataset by many folds and helps in reducing overfitting (10).

IMPLEMENTATION AND EXPERIMENTS

Implementation

Our python code uses the Keras package (16) and makes use of the Theano Library. Keras can leverage graphical processing units to accelerate the deep learning algorithms. We trained our CNN architecture on an NVIDIA Quadro M6000 GPU card. Training time was approximately 2 h.

Experimentation

We utilize the entire LUNA dataset ($n = 888$ patients), with 70% in training our model and 30% in test. In the training set, we balance our dataset for nodule states and non-nodule states. As shown in **Table 1**, for any sampling of states selected, approximately 5% are nodule states. Early on, the imbalance caused our model to bias significantly toward detecting non-nodule states given that those are the majority of states. The balanced dataset contains a total of 2,296 states, with 1,148 nodule states and 1,148 non-nodule states. It was created by retrieving nodule states from every patient with a nodule and random non-nodule states from all patients. For every epoch during the training, 20% of the training set is separated for cross-validation.

For our model, the sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) were computed as follows:

Sensitivity or true positive rate:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

Specificity or true negative rate:

$$\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{TN+FP}}$$

TABLE 1 | The number of patients and nodules they carry for nodule versus non-nodule groups.

	# of patients	# of states	# of nodules
Nodules	590	15,616	1,148
Non-nodules	298	7,107	0

Accuracy:

$$\frac{\mathrm{TP+TN}}{\mathrm{TP+FN+TN+FP}}$$

PPV:

$$\mathrm{PPV}=\frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

NPV:

$$\mathrm{NPV}=\frac{\mathrm{TN}}{\mathrm{TN+FN}}$$

where TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively.

RESULTS

As shown in **Figures 4** and **5**, for both loss and accuracy we observed a steady improvement. In **Figure 4**, showing the loss value over time, or epochs, there is a steady decline to approximately zero. A similar pattern holds with accuracy, in **Figure 5**, but with the steady increase to a value of one, meaning perfect score. Both graphs were generated from training on 70% of the dataset (1,607 states) and cross-validating on 20% of that (321 states). As observed in both graphs, the model is "learning", however there still remains considerable volatility as shown by the validation curves.

The conclusive results from the training and testing for our model is detailed in **Table 2** The test sample size was 30% of the dataset (668 states).

The testing results listed in **Table 2** are based on a cutoff value of 0.5. Given our model is a binary classifier, this means that for any state that it predicts, the likelihood of nodule is at least 0.5. **Figure 6** illustrates how the sensitivity and specificity vary as functions of cutoff values for both training and testing results.

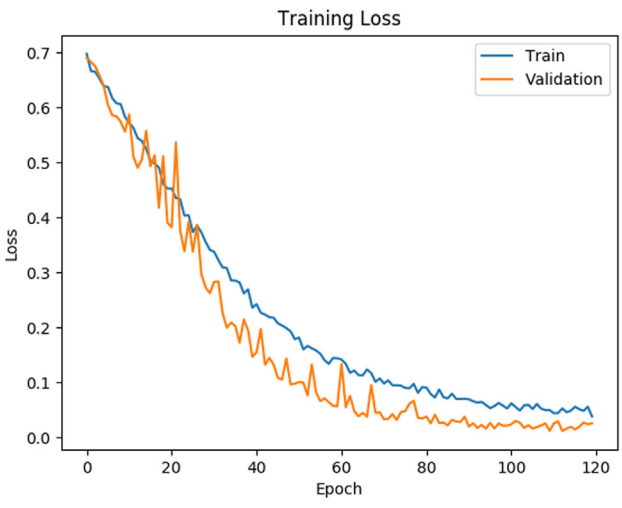

FIGURE 4 | Training and validation loss is shown on the training data for 120 epochs. Blue line corresponds to training loss and orange line corresponds to validation loss.

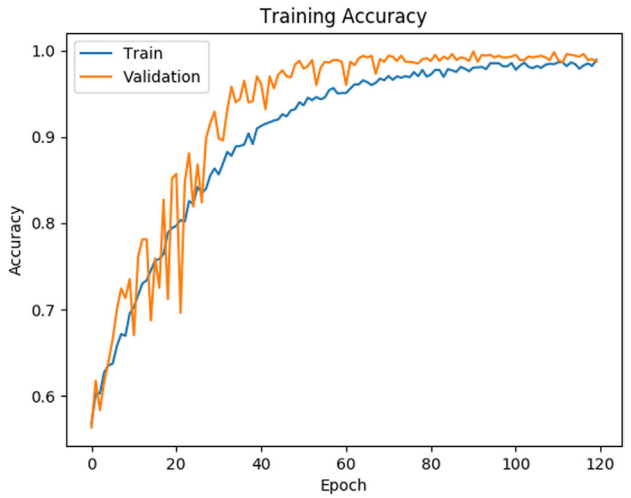

FIGURE 5 | Training and validation accuracy is shown for the training data for 120 epochs. Blue line corresponds to training accuracy and orange line corresponds to validation accuracy.

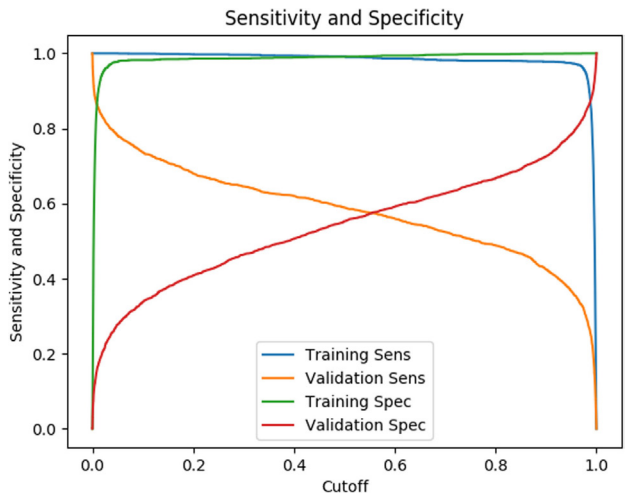

FIGURE 6 | Sensitivity and specificity as a function of cutoff, the likelihood a state has a nodule.

TABLE 2 | The sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) PPV results are listed for our reinforcement model from training and from testing.

	Accuracy	Sensitivity	Specificity	PPV	NPV
Training	99.1%	99.2%	99.1%	99.1%	99.2%
Test	64.4%	58.9%	55.3%	54.2.6%	60.0%

DISCUSSION

In this study, we present a robust non-invasive method to predict the presence of lung nodules, a common precursor to lung cancer, from lung CT scans using a RL method. A major advantage of this approach is that it allows to develop novel and unpredictable solutions to complex problems. From the results of our training using the LUNA dataset, we were able to achieve superb sensitivity, specificity, accuracy, PPV, and NPV (all greater than 99%). While the metrics for the testing dataset were lower, they were consistent. In both data size and number of trials, we achieved similar results. This consistency suggests that our research approach of using RL with non-pre-processed data is reproducible. Moreover, given the nature of RL, the model will only continue to improve with time and more data.

The way in which RL algorithms continue to improve depends not only on the quality of the dataset, but also more importantly its size. In the training of the AlphaGo, it was trained on master-level human players, instead of picking up the best strategies to win from scratch (8). In addition, the RL algorithm learned through more than 30 million human-on-human games. Factoring in hardware, AlphaGo required $25 million in computer hardware (17), it was trained on master-level human players (8).

Although the tasks of playing a game of Go is very different from detecting lung nodules, an inference we can draw is that reinforcement learning algorithms, such as AlphaGo, require substantial data to train. Given the original dataset's small size, there is an inherent difficulty in capturing the huge variability and structural differences in the lung volumes of human beings. With only 888 CT scans and approximately 1,148 nodule states in our dataset, with 70% of that being used for training, the lesson we have learned is that our model needs a significant amount of more data. This is evidenced by the tremendous amount of data and hardware needed to train AlphaGo to reach super human performance.

It is worth noting that AlphaGo's performance is based on how well it performed against human players. Similarly, our model performance is based on how well it performed against at least three radiologists in detecting lung nodules. As described by Armato et al. (9) how a given lesion was classified as a nodule was determined by a consensus of at least three of the four radiologists. A significant variability is observed when comparing the number of lesions classified as a nodule by one radiologist versus at least three radiologists. For the lesions identified in all the scans, 928 lesions were classified as nodules ≥3 mm by all four radiologists and 2,669 lesions were classified as a nodule ≥3 mm by at least one radiologist. This means for nodules ≥3 mm, the false discovery rate for a given individual radiologist is 65.2% (9). In contrast, despite the overfitting, our model classification yields a false discovery rate of 44.7% on the validation dataset, which is an improvement compared to an individual radiologist.

Given the very high training results, the question of overfitting arises. With a small dataset, the underlying probability distribution of lung nodules is not sufficient to create a fully generalizable model, especially given it is based on RL. As with most parametric tests, a fundamental assumption of samples is that they adequately capture the variance of the population they represent. With small datasets, depending on the variable, a random subset of the data may not adequately capture the variance of the overall dataset. With the LUNA dataset, this is particularly an issue given the fact that it is very high dimensional and our model requires significantly more data to capture the true variance of its countless variables. Most CT image datasets comprise of thousands of images, as compared to the millions of games in AlphaGo, and thus the comparison is not quite the same. We employed dropout and

data augmentation to increase the generalizability of our model in response to the overfitting. Together these two approaches have minimally dampened the effect. An alternate approach we also experimented with was to reduce the network size, however, this approach resulted in significant volatility in the training and validation results. Regardless of the overfitting, the performance on the validation data set indicates that our model achieves enough generalization to compete with a human radiologist and could serve as a second reader.

A strength of our research approach is the lack of pre-processing. It is known that medical imaging, including CT images, can be very heterogeneous. From the number of image slices, scanning machine used, and scanning parameters used, the image data for each patient is very disparate. A significant negative byproduct of this heterogeneity is the astronomical number of insignificant features generated that are unrelated to one's outcome of interest, such as the presence of lung nodules. For a machine learning algorithm to contend with this either the data size has to exponentially increase or many of the insignificant features have to be pre-processed out by filtering for only the relevant features. The former option of increasing the dataset is impractical, as the LUNA dataset is already one of the largest and most comprehensive image datasets. Hence, most, if not all, approaches in the current literature on CAD systems for lung nodule detection take the second option of pre-processing. From using various filters, masks, and general pre-processing tools, these methods heavily curate and alter the raw medical image data. As a result, this can create an infinite number of variations of the original dataset, and such a subjective practice makes it very difficult to reproduce any of the experimental results. We choose to use data without pre-processing to ensure that our results are reproducible.

Our work highlights the promise of using RL for lung nodule detection. There are several practical applications of this model, one of which is to serve as a second opinion or learning system for radiologists and trainees in identifying lung nodules. A strong appeal of using a RL approach is that the model is always in a learning state. With every new patient, the model expands its learning by factoring in the new information and building upon its probabilistic memory of historical information from previous patients. This phenomenon is what allowed the artificial intelligence model AlphaGo to keep improving after each match, eventually beating each player after several matches, including the reigning world champion. Likewise, we expect that our model will continue to improve as it observes more and more cases.

AUTHOR CONTRIBUTIONS

IA and GH: carried out primary experiments of project. GG, MK, and XM: provided guidance on methodology and overall project. YL, WM, and BN: provided lab and technical support. JD: generated research ideas, provided guidance on methodology and overall project, and reviewed manuscript.

FUNDING

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB022589.

REFERENCES

1. Moyer VA; U.S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med* (2014) 160(5):330–8. doi:10.7326/M13-2771
2. *CDC – Lung Cancer*. Available from: https://www.cdc.gov/cancer/lung/index.htm (Accessed: January 2, 2018).
3. *Cancer Moonshot*. National Cancer Institute. Available from: https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative (Accessed: December 6, 2017).
4. Swensen SJ, Jett JR, Hartman TE, Midthun DE, Mandrekar SJ, Hillman SL, et al. CT screening for lung cancer: five-year prospective experience. *Radiology* (2005) 235(1):259–65. doi:10.1148/radiol.2351041662
5. Midthun DE. Early detection of lung cancer. *F1000Res* (2016) 5:Faculty of 1000 Ltd. doi:10.12688/f1000research.7313.1
6. Caroline C. Lung cancer screening with low dose CT. *Radiol Clin North Am* (2014) 52(1):27–46. doi:10.1016/j.rcl.2013.08.006
7. Saba L, Caddeo G, Mallarini G. Computer-aided detection of pulmonary nodules in computed tomography: analysis and review of the literature. *J Comput Assist Tomogr* (2007) 31(4):611–9. doi:10.1097/rct.0b013e31802e29bf
8. Gibney E. Self-taught AI is best yet at strategy game go. *Nat News* (2017) 550:16–7. doi:10.1038/nature.2017.22858
9. Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* (2011) 38(2):915–31. doi:10.1118/1.3528204
10. Akkus Z, Ali I, Sedlář J, Kline TL, Agrawal JP, Parney IF, et al. *Predicting 1p19q Chromosomal Deletion of Low-Grade Gliomas from MR Images Using Deep Learning*. (2016). Available from: https://arxiv.org/abs/1611.06939 (Accessed: August 15, 2017).
11. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature* (2015) 518(7540):529–33. doi:10.1038/nature14236
12. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. *J Artif Intell Res* (1996) 4:237–85.
13. van Otterlo M, Wiering M. *Reinforcement Learning and Markov Decision Processes. Reinforcement Learning*. Berlin, Heidelberg: Springer (2012). p. 3–42.
14. Christopher JC, Dayan P. Q-learning. *Mach Learn* (1992) 8(3–4):279–92. doi:10.1023/A:1022676722315
15. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in Neural Information Processing Systems 25*. New York: Curran Associates, Inc. (2012). p. 1097–105.
16. Chollet F. *Keras*. Github (2015). Available from: https://github.com/fchollet/keras (Accessed: January 2, 2016).
17. Sutton RS. Introduction: the challenge of reinforcement learning. *Mach Learn* (1992) 8(3–4):225–7. doi:10.1023/A:1022620604568

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

PERSPECTIVE
published: 17 April 2018
doi: 10.3389/fonc.2018.00110

Machine Learning in Radiation Oncology: Opportunities, Requirements, and Needs

Mary Feng, Gilmer Valdes, Nayha Dixit and Timothy D. Solberg*

Department of Radiation Oncology, University of California San Francisco, San Francisco, CA, United States

Machine learning (ML) has the potential to revolutionize the field of radiation oncology, but there is much work to be done. In this article, we approach the radiotherapy process from a workflow perspective, identifying specific areas where a data-centric approach using ML could improve the quality and efficiency of patient care. We highlight areas where ML has already been used, and identify areas where we should invest additional resources. We believe that this article can serve as a guide for both clinicians and researchers to start discussing issues that must be addressed in a timely manner.

OPEN ACCESS

Edited by:
Jun Deng,
Yale University,
United States

Reviewed by:
Timothy Showalter,
University of Virginia,
United States
Maria Chan,
Memorial Sloan Kettering Cancer Center, United States

***Correspondence:**
Mary Feng
mary.feng@ucsf.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *22 February 2018*
Accepted: *29 March 2018*
Published: *17 April 2018*

Citation:
Feng M, Valdes G, Dixit N and Solberg TD (2018) Machine Learning in Radiation Oncology: Opportunities, Requirements, and Needs. Front. Oncol. 8:110. doi: 10.3389/fonc.2018.00110

Keywords: machine learning, radiation oncology, big data, predictive models, process improvement

INTRODUCTION

The expanding collection and sharing of data, increases in computational power, and perhaps most significantly, advances in machine learning (ML) and artificial intelligence, are rapidly transforming society, and offer the potential for similar transformation within health care. Ongoing advances in ML and big data analytics have spurred numerous efforts in precision oncology (1, 2), and the field of radiation oncology is uniquely poised to benefit from prudent application of such techniques. Radiation oncology has many specific challenges, however, ranging from unique datasets [e.g., 4DCT, CBCT, dose, structures, setup, and quality assurance (QA) information], limited clinical outcomes data, variation in dose and fraction schedules comprising standard of care, interaction of radiation and chemotherapy, limited access to genomics data, and other complexities. The historical reliance on empirical approaches such as the linear-quadratic model further influences clinical practice (3). Furthermore, the current ML hype is largely a result of success in a few, very specific tasks, such as image classifications, games, and autonomous driving systems (4–6). It is critical that we understand this success depends as much on the nature of the task as on the nature of the algorithm and the availability and quality of data, and thus meaningful gains in our field may prove more challenging.

In this article, we review the radiotherapy process from a workflow perspective, identifying specific areas where ML could improve the quality and efficiency of the current caregiving process for patients treated with radiation therapy. We have divided radiotherapy into six serial stages that encompass the entirety of treatment: patient assessment, simulation, planning, QA, treatment delivery, and follow-up, **Figure 1**. In each of these areas, we have identified open questions, emerging techniques, and possible directions concerning all stakeholders: patients, oncologists, physicists, dosimetrists, therapists, and nurses. Each stage is accompanied by a systematic assessment of opportunities, expectations, applicability, and limitations of various ML algorithms. While the impact a data-centric approach can have on improving the quality of treatment for cancer patients is clear, utilizing such a method will require a cultural shift at both the professional and institutional levels. We believe that this article will serve as a guide for both clinicians and researchers on those problems that must be addressed in a timely manner.

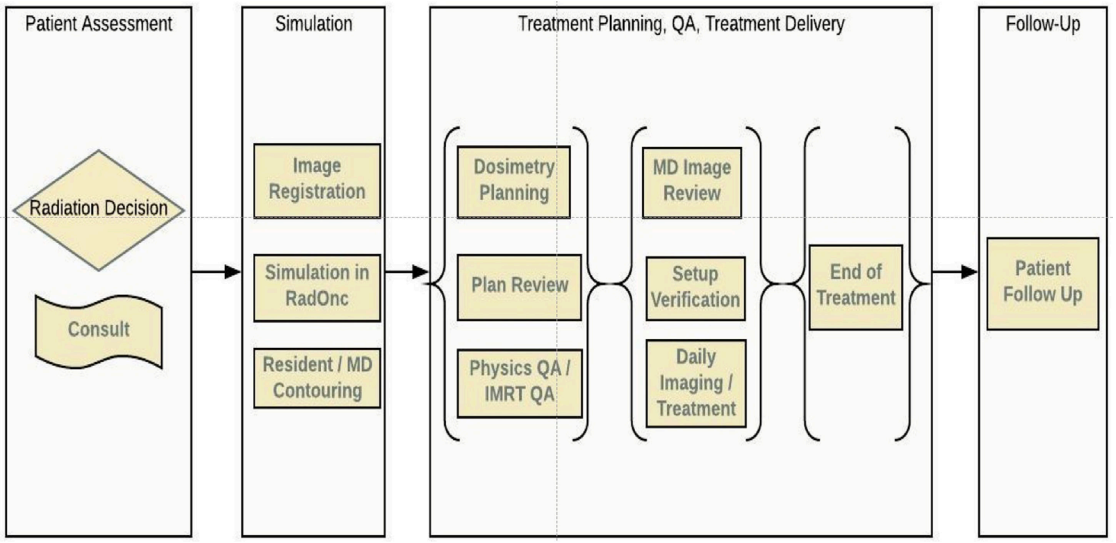

FIGURE 1 | Radiotherapy workflow, from consult to follow-up.

PATIENT ASSESSMENT

The radiation oncology process begins at the first consultation. During this time, the radiation oncologist and patient meet to discuss the clinical situation, including the risks and benefits of treatment and the patient's goals of care, to determine a treatment strategy. Useful information to assess the potential benefit of treatment includes tumor stage, mutational status or gene signatures (e.g., MGMT, Oncotype score), viral status (e.g., HPV), prior and current therapies, margin status if post-resection, ability to tolerate multimodality therapy, and overall performance status. Balanced with this are parameters that impact potential risk and tolerability of treatment including age, comorbidities, functional status, functioning of important organs, proximity between tumor and critical normal tissues, supportive care network, and ability to cooperate with motion management. All of these are features can be used to build predictive models of treatment outcome and toxicity. These models, then, can be used to inform physicians and patients to manage expectations and guide trade-offs between risks and benefit.

Having an ontology to identify and categorize the information available at this stage is important for any successful application of any predictive model (7). By contrast, current predictive models utilizing tumor control or normal tissue complication probability are neither subdivided nor categorized according to current state of a patient within the treatment timeline (8–12). Rather, they make use of a predetermined set of features, collected by individual investigators, that may have previously shown correlation to a particular clinical outcome (8–12). As a result, physicians are limited by scare and siloed data, making it often necessary to make informed guesses rather than data-driven decisions.

Many opportunities can be found for predictive models at the stage of initial consult. Here are a few practical examples in which a data-centric approach could improve decision making at the time of consultation:

(1) You are asked to see an inpatient who has a painful cervical spine metastasis. She will be discharged to hospice. What information may be helpful to determine whether to recommend RT? On one hand, radiotherapy can palliate her pain. However, she may not live long enough to benefit from treatment, but will have the discomfort associated with transfers and positioning for simulation/treatment, acute esophagitis, and pain flare. Narcotic management may be best for her, but how would we know? Models which predict time to pain relief, risk of toxicity, and overall survival would help optimize decision making at end-of-life, maximizing quality of life for the patient, and delivering high-value care (13).
(2) A patient with intermediate-risk prostate cancer is referred to discuss therapy. Treatment options for could include fractionated external beam radiotherapy, stereotactic body radiotherapy, brachytherapy, surgery, and other non-radiation approaches, without or without androgen-deprivation therapies. Shared decision making is crucial in this situation, since each type of therapy has inherent trade-offs with different side-effect profiles, which drives choice of therapy. A clinical support tool showing the balance between efficacy and side effects based on pre-treatment function and choice of therapy would be helpful for physicians and patients in shared decision making.
(3) A patient with hepatitis C cirrhosis and a single hepatocellular carcinoma is referred to consider treatment options. To determine whether to recommend SBRT over other treatments such as radiofrequency ablation or transarterial chemoembolization, information on liver function, suitability for anesthesia, and proximity to bowel, heart, gall bladder, central biliary tree is needed. Comprehensive ways

to integrate tumor control and toxicity predictions from all treatment modalities would help the physician and patient to manage expectations and decide on a course of therapy. Within radiation oncology, there has been some work done to model individual radiation sensitivity to individualize and adapt therapy, though there is still much opportunity for richer predictive modeling using ML (14).

(4) A patient with early stage left-sided breast cancer had a lumpectomy with negative margins and comes to discuss adjuvant radiotherapy. How would you decide whether she would benefit from deep inspiration breath-hold or intensity-modulated radiotherapy rather than 3D conformal RT? Would proton therapy be beneficial for this patient? Since the complexity and cost is higher with more advanced technology, models to predict who would benefit would be helpful for technology selection and resource allocation (11, 15–18).

The delivery of models that could help with these scenarios will require a cultural shift in our profession toward standardization and collaboration. In this regard, new collaborative projects have begun in recent years, though participation is not yet widespread (12, 19, 20). In addition, while recently published task reports have aimed to standardize nomenclature in radiotherapy (21), it is equally important to develop standards for data collection. Due in part to the small datasets typically encountered in radiotherapy, the choice of algorithm in a specific application can produce differences of up to 32% in predicted outcome (22). It is also important to understand the goal of any modeling effort. If the goal is to assist physicians and patients reach the best decision, then a balance between interpretability of the results and accurate predictions is needed (23, 24). In this case, logistic regressions or decision trees are equally effective (23, 24). If accuracy is favored over interpretability, then tree base methods such as random forests or gradient boosting, and Support Vector Machines with kernel methods, consistently win most modeling competitions when structured data are analyzed (such as the type of data described above) (25, 26).

SIMULATION

Once a physician and patient have decided to proceed with radiation therapy, the physician will place robust instructions for a Simulation, which is then scheduled. The order for simulation includes details about immobilization, scan range, treatment site, and other specifics necessary to complete the procedure appropriately. Patient preparation for simulation could include fiducial placement, fasting or bladder/rectal filling instructions, or kidney function testing for IV contrast. Special instructions are given for patients with a cardiac device or who are pregnant, and lift help or a translator is requested if necessary.

In most cases, a Simulation is scheduled after appropriate CT orders have been placed in the electronic medical record. Following completion and review of the CT simulation, the scan is exported to a planning system for the physician to contour tumor volumes and organs at risk (OARs). Sometimes a MR, PET and/or pre-operative scan is registered. With the OARs and tumor volumes contoured, the dosimetrists then begin designing a treatment plan based on specific physician instructions.

A good CT simulation is critical to the success of all subsequent processes, to achieve an accurate, high quality, robust, and deliverable plan for a patient. It is not uncommon that deficiencies at the time of CT simulation result in a need for a patient to return for a repeat CT, including insufficient scan range, incorrect IV contrast protocols, suboptimal immobilization, incorrect bladder/rectal filling, artifacts from internal hardware or those caused by the 4DCT process, lack of breath-hold reproducibility, and so on. Thus, focusing on the simulation, in particular, there are many questions that could be answered through ML algorithms to aid in decision making and overall workflow efficiency:

1. Will this patient benefit from IV contrast?
2. Will this patient be compliant with immobilization and motion management technique (e.g., compression or breath hold)?
3. Considering breathing patterns and other issues, will a 4DCT be beneficial for this patient?
4. Will this patient be able to tolerate the duration of the intended treatment (AP/PA vs. SBRT) and IGRT method (CBCT vs. kV-kV Orthogonal)?
5. Will this patient's anatomy allow for standard immobilization for simulation and treatment?

Simulation is an area where the community has focused little effort on ML, with early work confined to predicting tumor motion (27–30). Additional emphasis, from both academic institutions and industry, can be expected in the future.

TREATMENT PLANNING

The planning process starts by delineating both the target(s) and the OARs. While a number of commercial auto-segmentation algorithms exist, the underlying technology relies on an atlas-based strategy rather than utilizing ML. The performance of atlas-based segmentation tools depend highly on the type of structure, showing better results for high-contrast organs (e.g., lung) while struggling with soft tissue organs (e.g., pancreas) (31). By contrast, recent advances in computer vision, specifically around deep learning (6, 32), are particularly well suited for auto-segmentation (33–35). In deep learning, the algorithm is tasked to design the best features (higher order features) from the raw data as well to produce the classifiers (6). This is particularly important when human experts are unable to design proper features or quantify a given process, as in computer vision problems. An important limitation of the application of deep learning to segmentation is the limited size of the datasets available in radiation oncology. Because the algorithm is tasked to find the features as well as the classifier, deep learning models contain millions of parameters, and thus require more data than traditional ML algorithms. In applications in which deep learning has been successfully applied, the models have been trained with tens of thousands observations (4, 36). Although there are techniques to prevent overfitting when the number of parameters is larger than the number of observation points (transfer learning,

dropout, early stopping), it remains to be demonstrated whether these algorithms can generalize to datasets on the order of a few hundred in size, even when the techniques mentioned above are used. In our opinion, the effective application of deep learning to segmentation requires training and validation on datasets across multiple institutions and multiple scanners.

Once the target volumes and OARs have been delineated, the planning process continues by (1) setting dosimetric goals for targets and normal tissues; (2) selecting an appropriate treatment technique (e.g., 3D, fixed beam IMRT, VMAT, protons); (3) iteratively modifying the beams/weights/etc., until the planning goals have been achieved; (4) evaluating and approving the plan. It is in this last step where most ML applications have been focused (37–43). While there techniques are typically referred to as knowledge-based planning (KBP), it is important to highlight that both current academic research and available commercial products are limited to predicting dose–volume histograms (DVHs) within accepted ranges. Several authors have shown the value that DVH prediction has in improving population based treatment plan quality and in the detection of outliers (44–47).

Similar gains in steps 1–3 highlighted above would be equally important. For example, while KBP can predict DVHs, the intrinsic trade-offs between dosimetric indices that must be considered in step 1 are not currently predicted. A more recent commercial product, Quick Match (Siris Medical, Redwood City, CA, USA), uses gradient boosting (the most accurate algorithm on expectation when structured data are available) to explore predictions in dosimetric trade-offs (17). This application, which is similar to a treatment planning Pareto solution but obtained before the treatment planning process, can facilitate communication between dosimetrist and physicians, establish individualized and achievable goals, and help physicians and patients decide the course of plan before embarking on the treatment planning process. In addition, by allowing the exploration of intrinsic trade-offs, it can also help to choose an optimal technique (e.g., photon vs. protons).

Once the dosimetric goals have been established and the technique chosen, automatic plan generation is also possible. Attempts have been made to solve various aspects of this problem, for instance, predicting the best beam orientations (48, 49). The larger task of automated treatment planning, however, is well suited for reinforcement learning. In this technique, widely used in games, self-driving cars, and other popular-culture applications, an algorithm learns to navigate a set of rules, given some constraints, by self-correcting its decisions. For example, one could use fundamental laws of radiation interaction to achieve certain dosimetric constraints. Essentially, the algorithm will take a decision (for instance, increase the weight of a given constraint) and learn from the simulator (the treatment planning system) whether the decision resulted in the right direction. Common to successful applications of reinforcement learning is the ability to generate synthetic data using a simulator (e.g., games). This technique, successfully used by Google Brain to develop an algorithm capable of beating a Go world champion (5), could provide performance at the level of our best dosimetrists if properly implemented. One challenge of achieving automatic planning using reinforcement learning lies in the close integration that this research endeavor will need with robust treatment planning systems. Therefore, it seems likely that an industry/academic partnership is best suited to achieve this goal. Summarizing then, in the future, we envision the planning process to happen fully automatically, from contouring to plan creation, with the human experts (dosimetrists, physicists, and physicians) evaluating, supervising, and providing QA to the given results.

QA AND TREATMENT DELIVERY

A number of aspects of a radiotherapy QA program, specifically in error detection and prevention, treatment machine QA, and time series analyses, are well suited to the application of ML (50–53). Li and Chan developed an application to predict the performance of linear accelerators over time (51). Valdes et al. developed ML applications to predict IMRT QA passing rates (52, 53) and to automatically detect problems with the Linac imaging system (50). Carlson et al. developed a ML approach to predict multi-leaf collimator positional errors (54). El Naqa developed system to detect anomalies in QA data (55). Finally, Ford et al. have developed a tool to quantify the value of quality control checks in radiation oncology (56). The ability of these algorithms to automatically detect outliers allows physicists to focus attention on those aspects of a process most likely to impact our patient care, as recommended in Task Group 100 (57).

Other important applications of ML include predicting planning deviations from the initial intentions and predicting the need for re-planning. Guidi et al. developed a ML-based tool to predict when head and neck patients treated with photons need re-planning (58). In a similar fashion, Tseng et al. used three deep neural networks to predict the need for treatment adaptation for lung patients (59, 60). Varfalvy et al. used relative gamma analysis and hidden Markov models to categorize patients based on deviations from the initial treatment plan to identify patients in need of re-planning. The need to predict proton patients who would benefit from a re-plan is even more relevant, though no publications exist in this setting to date. Deciding on the best algorithm for QA applications is critical for accurately predicting outcomes. While it is clear that each of the applications described above are important and useful, all remain within the domain of research and have not been made available commercially.

FOLLOW-UP

Machine learning also has the potential to change the way radiation oncologists follow patients treated with definitive therapy. Following surgery, the tumor may disappear on imaging, and tumor markers may quickly normalize. By contrast, the evolution of imaging changes (loss of enhancement, PET avidity, or diffusion restriction; stability or decrease in size) and response of tumor markers is gradual following radiotherapy. These features are monitored regularly over time, with qualitative changes complemented with clinical experience providing indication of therapeutic efficacy. Clearly, better models based on early assessment are needed to predict outcome, in time for treatment intensification with additional RT, early addition of systemic therapy, or application of a different treatment modality. In this regard,

early work in the area of radiomics seems promising. In radiomics, quantitative features, including those based on size and shape, image intensity, texture, relationships between voxels, and fractal characteristics, are extracted to characterize an image. ML algorithms can then be deployed to correlate the image-based features with biological observations or clinical outcomes (59–64). The limited reproducibility of imaging systems both within and across institutions remains a significant challenge for radiomics (65, 66). And while the application of deep learning to image quantification has produced stellar results in other areas (67), it is important to understand that these techniques required thousands of data points even when transfer learning was used, which can prove challenging in radiation oncology, where datasets are limited.

CONCLUSION

Machine learning is poised to impact the profession of radiation oncology, from patient consult to follow-up. While the excitement around ML and big data is well justified, many challenges remain, a number of which we have tried to describe above. There are also several broad challenges we will have to address as a field. The first is the creation and curation of large datasets. Although it is highly unlikely that robust models can be built with data from a single institution alone, the need to develop data sharing agreements can be a significant barrier to the development of these models. One potential solution to the challenges associated with multi-institutional data sharing is the use of distributed learning; the group at Maastricht University led by Philippe Lambin has been pioneering this approach (68, 69). Standardization of the data collection process is also essential for training models using datasets from multiple institutions. In addition, it is important to highlight that distributed learning and transfer learning are part of the larger discipline of ML and to maximize learning from all centers while customizing the solution to each, mathematical guarantees and constraints are necessary to ensure algorithms do not "forget" previous seen datasets (70). Tailoring these algorithms to radiation oncology needs will also be an active research area in the future. Quality of data is also of paramount importance as no ML algorithm today can fix problems contained within the training data. In this regard, interpretability of algorithms used (e.g., ability for humans experts to understand reasons behind a prediction) will play an important role to avoid preventable errors (23).

Finally, training of our workforce and updating our educational curriculums will be increasingly important. As with any algorithm that we use in radiation oncology today (e.g., dose calculation, deformable registration), ML algorithms will need commissioning and QA. Clinicians will need to learn to interpret and understand the limitations of any results. The field of radiation oncology is highly algorithmic and data-centric, and while the road ahead is filled with potholes, the destination holds tremendous promise.

AUTHOR CONTRIBUTIONS

All authors were involved with the conception and design, manuscript writing, and final approval of the manuscript.

REFERENCES

1. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. *J Clin Oncol* (2013) 31:1803–5. doi:10.1200/JCO.2013.49.4799
2. Biankin AV, Piantadosi S, Hollingsworth SJ. Patient-centric trials for therapeutic development in precision oncology. *Nature* (2015) 526:361. doi:10.1038/nature15819
3. Bentzen SM, Constine LS, Deasy JO, Eisbruch A, Jackson A, Marks LB, et al. Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *Int J Radiat Oncol Biol Phys* (2010) 76:S3–9. doi:10.1016/j.ijrobp.2009.09.040
4. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. Lake Tahoe (2012). p. 1097–105.
5. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of go with deep neural networks and tree search. *Nature* (2016) 529:484–9. doi:10.1038/nature16961
6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* (2015) 521:436–44. doi:10.1038/nature14539
7. Miller AA. *Developing an Ontology for Radiation Oncology, Master of Information and Communication Technology - Research thesis, School of Information Systems and Technology, University of Wollongong* (2012). Available from: http://ro.uow.edu.au/theses/3744
8. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB II. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Phys Med Biol* (2016) 61:6105. doi:10.1088/0031-9155/61/16/6105
9. Naqa IE, Deasy JO, Mu Y, Huang E, Hope AJ, Lindsay PE, et al. Datamining approaches for modeling tumor control probability. *Acta Oncol* (2010) 49:1363–73. doi:10.3109/02841861003649224
10. El Naqa I, Bradley J, Blanco AI, Lindsay PE, Vicic M, Hope A, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. *Int J Radiat Oncol Biol Phys* (2006) 64:1275–86. doi:10.1016/j.ijrobp.2005.11.022
11. Langendijk JA, Lambin P, De Ruysscher D, Widder J, Bos M, Verheij M. Selection of patients for radiotherapy with protons aiming at reduction of side effects: the model-based approach. *Radiother Oncol* (2013) 107:267–73. doi:10.1016/j.radonc.2013.05.007
12. Lambin P, Roelofs E, Reymen B, et al. 'Rapid learning health care in oncology' – an approach towards decision support systems enabling customised radiotherapy. *Radiother Oncol* (2013) 109:159–64. doi:10.1016/j.radonc.2013.07.007
13. Kress MA, Jensen RE, Tsai HT, Lobo T, Satinsky A, Potosky AL. Radiation therapy at the end of life: a population-based study examining palliative treatment intensity. *Radiat Oncol* (2015) 10:15. doi:10.1186/s13014-014-0305-4
14. Feng M, Suresh K, Schipper MJ, Bazzi L, Ben-Josef E, Matuszak MM, et al. Individualized adaptive stereotactic body radiotherapy for liver tumors in patients at high risk for liver damage: a phase 2 clinical trial. *JAMA Oncol* (2018) 4:40–7. doi:10.1001/jamaoncol.2017.2303
15. Blanchard P, Wong AJ, Gunn GB, Garden AS, Mohamed ASR, Rosenthal DI, et al. Toward a model-based patient selection strategy for proton therapy: external validation of photon-derived normal tissue complication probability models in a head and neck proton therapy cohort. *Radiother Oncol* (2016) 121:381–6. doi:10.1016/j.radonc.2016.08.022
16. Hall DC, Trofimov AV, Winey BA, Liebsch NJ, Paganetti H. Predicting patient-specific dosimetric benefits of proton therapy for skull-base tumors using a geometric knowledge-based method. *Int J Radiat Oncol Biol Phys* (2017) 97:1087–94. doi:10.1016/j.ijrobp.2017.01.236
17. Valdes G, Simone CB II, Chen J, Lin A, Yom SS, Pattison AJ, et al. Clinical decision support of radiotherapy treatment planning: a data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiother Oncol* (2017) 125(3):392–7. doi:10.1016/j.radonc.2017.10.014
18. Everett-Thomas R, Valdes B, Valdes GR, Shekhter I, Fitzpatrick M, Rosen LF, et al. Using simulation technology to identify gaps between education and

practice among new graduate nurses. *J Contin Educ Nurs* (2015) 46:34–40. doi:10.3928/00220124-20141122-01
19. Moran JM, Feng M, Benedetti LA, Marsh R, Griffith KA, Matuszak MM, et al. Development of a model web-based system to support a statewide quality consortium in radiation oncology. *Pract Radiat Oncol* (2017) 7:e205–13. doi:10.1016/j.prro.2016.10.002
20. Bowers MR, McNutt TR, Wong JW, Phillips MH, Hendrickson KRG, Kwok P, et al. Oncospace consortium: a shared radiation oncology database system designed for personalized medicine and research. *Int J Radiat Oncol Biol Phys* (2015) 93:E385. doi:10.1016/j.ijrobp.2015.07.1529
21. Mayo CS, Moran JM, Bosch W, Xiao Y, McNutt T, Popple R, et al. AAPM task group 263: tackling standardization of nomenclature for radiation therapy. *Int J Radiat Oncol Biol Phys* (2015) 93:E383–4. doi:10.1016/j.ijrobp.2015.07.1525
22. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep* (2015) 5:13087. doi:10.1038/srep13087
23. Valdes G, Luna JM, Eaton E, Simone CB II, Ungar LH, Solberg TD. MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Sci Rep* (2016) 6:37854. doi:10.1038/srep37854
24. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predict-ing pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM (2015). p. 1721–30.
25. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh: ACM (2006). p. 161–8.
26. Fernández-Delgado M, Cernadas E, Barro S, Amorin D. Do we need hundreds of classifiers to solve real world classification problems. *J Mach Learn Res* (2014) 15:3133–81.
27. Ruan D, Keall P. Online prediction of respiratory motion: multidimensional processing with low-dimensional feature learning. *Phys Med Biol* (2010) 55:3011. doi:10.1088/0031-9155/55/11/002
28. Ruan D, Fessler JA, Balter JM. Mean position tracking of respiratory motion. *Med Phys* (2008) 35:782–92. doi:10.1118/1.2825616
29. Ruan D. Kernel density estimation-based real-time prediction for respiratory motion. *Phys Med Biol* (2010) 55:1311. doi:10.1088/0031-9155/55/5/004
30. Isaksson M, Jalden J, Murphy MJ. On using an adaptive neural network to predict lung tumor motion during respiration for radiotherapy applications. *Med Phys* (2005) 32:3801–9. doi:10.1118/1.2134958
31. Delpon G, Escande A, Ruef T, Darréon J, Fontaine J, Noblet C, et al. Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy. *Front Oncol* (2016) 6:178. doi:10.3389/fonc.2016.00178
32. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys* (2017) 44:6377–89. doi:10.1002/mp.12602
33. Yang X, Wu N, Cheng G, Zhou Z, Yu DS, Beitler JJ, et al. Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal MRI study in head-and-neck radiation therapy. *Int J Radiat Oncol Biol Phys* (2014) 90:1225–33. doi:10.1016/j.ijrobp.2014.08.350
34. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* (2017) 44:547–57. doi:10.1002/mp.12045
35. Dai W, Doyle J, Liang X, Zhang H, Dong N, Li Y, et al. Scan: Structure Correcting Adversarial Network for Chest x-Rays Organ Segmentation. *arXiv preprint arXiv:1703.08770*. (2017).
36. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging* (2016) 35:1299–312. doi:10.1109/TMI.2016.2535302
37. Boutilier JJ, Craig T, Sharpe MB, Chan TC. Sample size requirements for knowledge-based treatment planning. *Med Phys* (2016) 43:1212–21. doi:10.1118/1.4941363
38. Schreibmann E, Fox T. Prior-knowledge treatment planning for volumetric arc therapy using feature-based database mining. *J Appl Clin Med Phys* (2014) 15:4596. doi:10.1120/jacmp.v15i2.4596
39. Chanyavanich V, Das SK, Lee WR, Lo JY. Knowledge-based IMRT treatment planning for prostate cancer. *Med Phys* (2011) 38:2515–22. doi:10.1118/1.3574874
40. Good D, Lo J, Lee WR, Wu QJ, Yin FF, Das SK. A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning. *Int J Radiat Oncol Biol Phys* (2013) 87:176–81. doi:10.1016/j.ijrobp.2013.03.015
41. Tol JP, Delaney AR, Dahele M, Slotman BJ, Verbakel WF. Evaluation of a knowledge-based planning solution for head and neck cancer. *Int J Radiat Oncol Biol Phys* (2015) 91:612–20. doi:10.1016/j.ijrobp.2014.11.014
42. Chang ATY, Hung AWM, Cheung FWK, Lee MCH, Chan OSH, Philips H, et al. Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy. *Int J Radiat Oncol Biol Phys* (2016) 95:981–90. doi:10.1016/j.ijrobp.2016.02.017
43. Yang Y, Xing L. Clinical knowledge-based inverse treatment planning. *Phys Med Biol* (2004) 49:5101–17. doi:10.1088/0031-9155/49/22/006
44. Shiraishi S, Tan J, Olsen LA, Moore KL. Knowledge-based prediction of plan quality metrics in intracranial stereotactic radiosurgery. *Med Phys* (2015) 42:908–17. doi:10.1118/1.4906183
45. Moore KL, Brame RS, Low DA, Mutic S. Experience-based quality control of clinical intensity-modulated radiotherapy planning. *Int J Radiat Oncol Biol Phys* (2011) 81:545–51. doi:10.1016/j.ijrobp.2010.11.030
46. Ahmed S, Nelms B, Gintz D, Caudell J, Zhang G, Moros EG, et al. A method for a priori estimation of best feasible DVH for organs-at-risk: validation for head and neck VMAT planning. *Med Phys* (2017) 44:5486–97. doi:10.1002/mp.12500
47. Fried DV, Chera BS, Das SK. Assessment of PlanIQ feasibility DVH for head and neck treatment planning. *J Appl Clin Med Phys* (2017) 18:245–50. doi:10.1002/acm2.12165
48. Rowbottom CG, Webb S, Oldham M. Beam-orientation customization using an artificial neural network. *Phys Med Biol* (1999) 44:2251. doi:10.1088/0031-9155/44/9/312
49. Llacer J, Li S, Agazaryan N, Promberger C, Solberg TD. Non-coplanar automatic beam orientation selection in cranial IMRT: a practical methodology. *Phys Med Biol* (2009) 54:1337. doi:10.1088/0031-9155/54/5/016
50. Valdes G, Morin O, Valenciaga Y, Kirby N, Pouliot J, Chuang C. Use of TrueBeam developer mode for imaging QA. *J Appl Clin Med Phys* (2015) 16:5363. doi:10.1120/jacmp.v16i4.5363
51. Li Q, Chan MF. Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study. *Ann N Y Acad Sci* (2017) 1387:84–94. doi:10.1111/nyas.13215
52. Valdes G, Scheuermann R, Hung CY, Olszanski A, Bellerive M, Solberg TD. A mathematical framework for virtual IMRT QA using machine learning. *Med Phys* (2016) 43:4323–34. doi:10.1118/1.4953835
53. Valdes G, Chan MF, Lim SB, Scheuermann R, Deasy JO, Solberg TD. IMRT QA using machine learning: a multi-institutional validation. *J Appl Clin Med Phys* (2017) 18(5):279–84. doi:10.1002/acm2.12161
54. Carlson JN, Park JM, Park SY, Park JI, Choi Y, Ye SJ. A machine learning approach to the accurate prediction of multi-leaf collimator positional errors. *Phys Med Biol* (2016) 61:2514. doi:10.1088/0031-9155/61/6/2514
55. El Naqa I. SU-E-J-69: an anomaly detector for radiotherapy quality assurance using machine learning. *Med Phys* (2011) 38:3458. doi:10.1118/1.3611837
56. Ford EC, Terezakis S, Souranis A, Harris K, Gay H, Mutic S. Quality control quantification (QCQ): a tool to measure the value of quality control checks in radiation oncology. *Int J Radiat Oncol Biol Phys* (2012) 84:e263–9. doi:10.1016/j.ijrobp.2012.04.036
57. Huq MS, Fraass BA, Dunscombe PB, Gibbons JP Jr, Ibbott GS, Mundt AJ, et al. The report of task group 100 of the AAPM: application of risk analysis methods to radiation therapy quality management. *Med Phys* (2016) 43:4209–62. doi:10.1118/1.4947547
58. Guidi G, Maffei N, Meduri B, D'Angelo E, Mistretta GM, Ceroni P, et al. A machine learning tool for re-planning and adaptive RT: a multicenter cohort investigation. *Phys Med* (2016) 32:1659–66. doi:10.1016/j.ejmp.2016.10.005
59. Tseng HH, Luo Y, Cui S, Chien JT, Ten Haken RK, Naqa IE. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med Phys* (2017) 44:6690–705. doi:10.1002/mp.12625
60. Varfalvy N, Piron O, Cyr MF, Dagnault A, Archambault L. Classification of changes occurring in lung patient during radiotherapy using relative γ analysis and hidden Markov models. *Med Phys* (2017) 44:5043–50. doi:10.1002/mp.12488

61. Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ. Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Sci Rep* (2017) 7:1648. doi:10.1038/s41598-017-01931-w
62. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep* (2017) 7:10353. doi:10.1038/s41598-017-10649-8
63. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep* (2017) 7:5467. doi:10.1038/s41598-017-05848-2
64. Cha KH, Hadjiiski L, Chan HP, Weizer AZ, Alva A, Cohan RH, et al. Bladder cancer treatment response assessment in CT using radiomics with deep-learning. *Sci Rep* (2017) 7:8738. doi:10.1038/s41598-017-09315-w
65. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48:441–6. doi:10.1016/j.ejca.2011.11.036
66. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* (2014) 5:4006. doi:10.1038/ncomms5006
67. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* (2016) 316:2402–10. doi:10.1001/jama.2016.17216
68. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital – a real life proof of concept. *Radiother Oncol* (2016) 121:459–67. doi:10.1016/j.radonc.2016.10.002
69. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. *Int J Radiat Oncol Biol Phys* (2017) 99:344–52. doi:10.1016/j.ijrobp.2017.04.021
70. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* (2011) 3:1–122. doi:10.1561/2200000016

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer MC declared a past collaboration with several of the authors GV, TS to the handling Editor.

frontiers in Oncology

PERSPECTIVE
published: 09 May 2018
doi: 10.3389/fonc.2018.00155

How Big Data, Comparative Effectiveness Research, and Rapid-Learning Health-Care Systems Can Transform Patient Care in Radiation Oncology

*Jason C. Sanders and Timothy N. Showalter**

Department of Radiation Oncology, University of Virginia School of Medicine, Charlottesville, VA, United States

Big data and comparative effectiveness research methodologies can be applied within the framework of a rapid-learning health-care system (RLHCS) to accelerate discovery and to help turn the dream of fully personalized medicine into a reality. We synthesize recent advances in genomics with trends in big data to provide a forward-looking perspective on the potential of new advances to usher in an era of personalized radiation therapy, with emphases on the power of RLHCS to accelerate discovery and the future of individualized radiation treatment planning.

Keywords: big data, radiation oncology, comparative effectiveness research, rapid-learning health care system, personalized radiation therapy

OPEN ACCESS

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
James Byunghoon Yu,
Yale University, United States
Andre Konski,
University of Pennsylvania, United States

***Correspondence:**
Timothy N. Showalter
tns3b@virginia.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *13 February 2018*
Accepted: *24 April 2018*
Published: *09 May 2018*

Citation:
Sanders JC and Showalter TN (2018) How Big Data, Comparative Effectiveness Research, and Rapid-Learning Health-Care Systems Can Transform Patient Care in Radiation Oncology. Front. Oncol. 8:155. doi: 10.3389/fonc.2018.00155

COMPARATIVE EFFECTIVENESS RESEARCH (CER) AND BIG DATA

The Committee on CER Prioritization was created by the Institute of Medicine in 2009. They defined CER as "a strategy that focuses on the practical comparison of two or more health intervention to discern what works best for which patients and populations" (1). In essence, the goal of CER is to help answer the question "which treatment will work best, in which patient, under what circumstances?" (2). Big Data refers to data sets that are so large that they cannot be analyzed directly by individuals or traditional processing software. Big Data Analytics (BDA) is a growing field with a multitude of methods that is being utilized in various sectors from business to medicine (3). The advent of the Electronic Medical Record (EMR) has resulted in the digitalization of massive data sets of medical information including: clinic encounters, laboratory values, imaging data sets and reports, pathology reports, patient outcomes, family history, genomic, and biological data, etc.

To help with the analysis of Big Data, the NIH has created the Big Data to Knowledge (BD2K) program which has invested over $200 million in grant awards to foster the development of methods and tools to analyze Big Data in biomedical research (4). Additionally, the BD2K program will move to make sure that biomedical Big Data is "Findable, Accessible, Interoperable, and Reusable" (4). Over the past decade, CER methodologies have become increasingly prevalent in radiation oncology research and there is much enthusiasm surrounding BDA.

RAPID-LEARNING HEALTH CARE SYSTEM (RLHCS) AND PERSONALIZED MEDICINE

The number of articles on Big Data in health care has increased exponentially from under 500 articles in 2005 to over 2500 articles in 2015 (5). As the amount of biomedical Big Data and our ability to analyze these data continues to advance, so will the implications and utilizations of the

information we are able to extract. One of the most important steps toward advancing our ability to analyze these Big Data for biomedical discovery is the creation of RLHCS, which will allow for the sharing of patient data between EMRs, ideally in real-time (6). An ideal RLHCS would take patient data that was routinely generated as part of standard patient care and compile that data into a large data system (6–8). This aggregate data would then be available for both BDA to accelerate identification of new hypotheses and CER to rapidly generate evidence through hypothesis-testing studies. Clinical data from patient records can be used readily to identify novel relationships among clinical factors and patient outcomes, or to evaluate treatment effectiveness in specific subgroups, that cannot be studied adequately in randomized, controlled trials. The extreme power of RLCHS, though, is even more exciting when one considers the possibility of adding biospecimens to accelerate discovery in genomics and proteomics. As RLHCS are created and their data sets are expanded, we will continue to identify specific genomic and proteomic data to help define cohorts and stratify patients into risk groups, treatment response groups, and potentially to help design highly tailored therapy regimens (9). In this sense, RLCHS would usher in a more fertile era for improving biomedical research than ever before. BDA and CER provide the research methodologies needed to rapidly generate evidence using RLHCS. It should be noted, however, that there are substantial practical obstacles that must be addressed to achieve the vision of RLHSC. These include patient concerns regarding privacy and security of sensitive information, interconnectivity among different health records, and regulatory barriers to the exchange of health information.

INTEGRATING A RLHCS WITH ONCOLOGY

The integration of CER, Big Data, and BDA is especially important in the field of Oncology where multiple groups are investing significant time and resources in efforts to expand the availability of data and advance the methods used to extract meaningful information from that data (4, 10–14). The American Society of Clinical Oncology started their own RLHCS, CancerLinQ, to overcome the lack of interoperability between EMRs and accomplish their goal of being able to "analyze and share data on every patient with cancer" (15). While the vision of RLCHS has not yet been fully achieved, the potential impact on society has stimulated enthusiasm toward this effort.

IMPLICATIONS FOR RADIATION ONCOLOGY

Patient Reported Outcomes (PROs)

Patient reported outcomes and quality-of-life (QoL) have become a major area of focus in health care overall, particularly in oncology. The availability of PROs within EMRs provides the foundation for a RLHCS that can be leveraged to expand insights into how cancer treatments impact patient QoL. By incorporating the PROs for massive numbers of patients, RLHCS will be able to identify small variations and subgroups of patients that might be missed in the smaller number of patients included in traditional randomized controlled trials. These PROs and QoL domains can then be incorporated into clinical decision-making to help guide both providers and patients (16). In doing this, PROs can act as a link between the objective clinical data and the subjective patient outcomes and experiences to help improve the overall care of the patient (17). One may also conceive of potential genomics-based determinants of QoL that could be identified using BDA if RLHCS include biospecimens linked to clinical data and PROs. Finally, surveillance of a RLHCS may also be performed to identify temporal trends in PROs to estimate outcomes after implementation of new technologies.

Dose Selection and Radiosensitivity

The use of tumor-specific genes and radiosensitivity to guided treatment decisions has already been established in human papilloma virus-associated squamous-cell carcinoma of the oropharynx (18). Numerous studies have looked at identifying genes that may have implications on tumor radiosensitivity or patient toxicity (19–22). The identification of these genes and their potential implications has led to the creation of the fields of radiogenetics and radiogenomics. Efforts are currently underway to generate meaningful gene assays that will help predict tumor response to radiation. Eschrich et al. created a 10-gene model to calculate a radiosensitivity index and applied this to patients with head-and-neck, rectal, and esophageal cancer to help stratify patients into either responders or non-responders with 80% sensitivity and 82% specificity (22). Similarly, Zhao et al. retrospectively created a 24-gene assay and applied this to risk matched patients who either received postoperative radiation or no radiation following prostatectomy. Patients with a high score on the gene index who received postoperative radiation were less likely to have distant metastasis at 10 years (23). As efforts to identify genes and gene assays that may be predictors of radiosensitivity continue to be validated, we will potentially be able to integrate these findings in dose selection and toxicity prediction for individual patients based on their native and tumor genetics. Scott and colleagues have recently described a genomics-based strategy for personalizing radiation therapy dose, which would support dose de-escalation for radiosensitive tumors (24). While the clinical implication of radiosensitivity assays are still developing, big data will be key to developing future assays rapidly, as well as incorporating the genomics tools into clinical decision-making. Big data provides opportunity to refine molecular signatures based upon real-world data and to merge genomic assay results with other clinical data elements to optimize predictive analytics. A RLHCS would provide the ideal substrate for levering big data and CER to accelerate genomics-based discovery to make precision radiation oncology a reality.

Personalized Treatment Recommendations

Radiation oncology is unique in that treatment plans for patients are often already technically and physically personalized due

to patient-specific variations in anatomy, tumor characteristics, and stage. Since a patient's treatment plan is usually based upon a CT scan in treatment position, radiation can be considered an inherently personalized form of medicine. However, treatment planning approaches and radiation doses are generally selected based upon class solution, with technical details such as beam arrangements and dose–volume constraints adherent to generalized rules. Multiple studies have already begun to look at how BDA methods such as machine learning and neural networks can be used to aid in dose optimization and toxicity prediction modeling in radiation oncology (17, 25–27), which could provide more optimal treatment plan alternatives for individual patients. As the data and technology behind RLHCS continues to progress, we will likely be able to utilize a full spectrum of patient-specific clinical factors, PROs, genomics, patient preference, and priorities, and a menu of treatment plan alternatives in order to optimize an individual patient's radiation therapy. In order to deliver high-quality, high impact insights into radiation oncology, it is important that large datasets include detailed technical.

CONCLUSION

Much of the excitement regarding big data has centered on potential for genomic discovery, high-level radiation treatment planning, and leveraging EMRs to identify associations among factors that may provide new insights into potential causal relationships that can be further studied to accelerate progress in cancer care. Although these are certainly promising areas for discovery, we most eagerly anticipate the power of big data to connect a broad range of characteristics to accelerate evidence generation and inform personalized decision-making. We envision the use of big data and CER methods to inform the individual decisions of patients and providers by synthesizing clinical and genomic data and querying a RLHCS for the latest data on effectiveness of treatment options in relevant subgroups of patients.

AUTHOR CONTRIBUTIONS

Both authors contributed to the development and editing of the manuscript and approved the final submitted version.

REFERENCES

1. Institute of Medicine of the National Academies. *Initial National Priorities for Comparative Effectiveness Research*. Washington, DC: Institute of Medicine of the National Academies (2009).
2. Greenfield S, Rich E. Welcome to the Journal of Comparative Effectiveness Research. *J Comp Eff Res* (2012) 1:1–3. doi:10.2217/cer.11.13
3. Sivarajah U, Kamal M, Irani Z, Weerakkody V. Critical analysis of big data challenges and analytical methods. *J Bus Res* (2017) 70:263–86. doi:10.1016/j.jbusres.2016.08.001
4. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc* (2014) 21:957–8. doi:10.1136/amiajnl-2014-002974
5. de la Torre Diez I, Cosgaya HM, Garcia-Zapirain B, Lopez-Coronado M. Big data in health: a literature review from the year 2005. *J Med Syst* (2016) 40:209. doi:10.1007/s10916-016-0565-7
6. Ginsburg GS, Kuderer NM. Comparative effectiveness research, genomics-enabled personalized medicine, and rapid learning health care: a common bond. *J Clin Oncol* (2012) 30:4233–42. doi:10.1200/JCO.2012.42.6114
7. Ginsburg GS, Staples J, Abernethy AP. Academic medical centers: ripe for rapid-learning personalized health care. *Sci Transl Med* (2011) 3:101cm27. doi:10.1126/scitranslmed.3002386
8. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-learning system for cancer care. *J Clin Oncol* (2010) 28:4268–74. doi:10.1200/JCO.2010.28.5478
9. Ramsey SD, Veenstra D, Tunis SR, Garrison L, Crowley JJ, Baker LH. How comparative effectiveness research can help advance 'personalized medicine' in cancer treatment. *Health Aff (Millwood)* (2011) 30:2259–68. doi:10.1377/hlthaff.2010.0637
10. Helft M. Can big data cure cancer? *Fortune* (2014) 170:70–4.
11. Williams AM, Liu Y, Regner KR, Jotterand F, Liu P, Liang M. Artificial intelligence, physiological genomics, and precision medicine. *Physiol Genomics* (2018) 50(4):237–43. doi:10.1152/physiolgenomics.00119.2017
12. Savage N. Big data versus the big C. *Sci Am* (2014) 311:S20–1. doi:10.1038/scientificamerican0714-S20
13. Shah A, Stewart AK, Kolacevski A, Michels D, Miller R. Building a rapid learning health care system for oncology: why CancerLinQ collects identifiable health information to achieve its vision. *J Clin Oncol* (2016) 34:756–63. doi:10.1200/JCO.2015.65.0598
14. Trifiletti DM, Showalter TN. Big data and comparative effectiveness research in radiation oncology: synergy and accelerated discovery. *Front Oncol* (2015) 5:274. doi:10.3389/fonc.2015.00274
15. American Society of Clinical Oncology. *Shaping the Future of Oncology: Envisioning Cancer Care in 2030, Outcomes of the ASCO Board of Directors Strategic Planning and Visioning Process, 2011-2012*. Alexandria, VA: American Society of Clinical Oncology (2012).
16. Sarin R. Big data V4 for integrating patient reported outcomes and quality-of-life indices in clinical practice. *J Cancer Res Ther* (2014) 10:453–5. doi:10.4103/0973-1482.142741
17. Kim KH, Lee S, Shim JB, Chang KH, Cao Y, Choi SW, et al. Predictive modelling analysis for development of a radiotherapy decision support system in prostate cancer: a preliminary study. *J Radiother Pract* (2017) 16:161–70. doi:10.1017/S1460396916000583
18. Chen AM, Felix C, Wang PC, Hsu S, Basehart V, Garst J, et al. Reduced-dose radiotherapy for human papillomavirus-associated squamous-cell carcinoma of the oropharynx: a single-arm, phase 2 study. *Lancet Oncol* (2017) 18:803–11. doi:10.1016/S1470-2045(17)30246-2
19. West CM, Barnett GC. Genetics and genomics of radiotherapy toxicity: towards prediction. *Genome Med* (2011) 3:52. doi:10.1186/gm268
20. Torres-Roca JF, Eschrich S, Zhao H, Bloom G, Sung J, McCarthy S, et al. Prediction of radiation sensitivity using a gene expression classifier. *Cancer Res* (2005) 65:7169–76. doi:10.1158/0008-5472.CAN-05-0656
21. Chistiakov DA, Voronova NV, Chistiakov PA. Genetic variations in DNA repair genes, radiosensitivity to cancer and susceptibility to acute tissue reactions in radiotherapy-treated cancer patients. *Acta Oncol* (2008) 47:809–24. doi:10.1080/02841860801885969
22. Eschrich SA, Pramana J, Zhang H, Zhao H, Boulware D, Lee JH, et al. A gene expression model of intrinsic tumor radiosensitivity: prediction of response and prognosis after chemoradiation. *Int J Radiat Oncol Biol Phys* (2009) 75:489–96. doi:10.1016/j.ijrobp.2009.06.014
23. Zhao SG, Chang SL, Spratt DE, Erho N, Yu M, Ashab HA-D, et al. Development and validation of a 24-gene predictor of response to postoperative radiotherapy in prostate cancer: a matched, retrospective analysis. *Lancet Oncol* (2016) 17:1612–20. doi:10.1016/S1470-2045(16)30491-0
24. Scott JG, Berglund A, Schell MJ, Mihaylov I, Fulp WJ, Yue B, et al. A genome-based model for adjusting radiotherapy dose (GARD): a retrospective, cohort-based study. *Lancet Oncol* (2017) 18:202–11. doi:10.1016/S1470-2045(16)30648-9

25. Kim KH, Lee S, Shim JB, Chang KH, Yang DS, Yoon WS, et al. Mining and toxicity prediction modeling system for a clinical decision support in radiation oncology: a preliminary study. *J Korean Phys Soc* (2017) 71:231–7. doi:10.3938/jkps.71.231
26. Arimura H, Nakamoto T. Applications of machine learning for radiation therapy. *Igaku Butsuri* (2016) 36:35–8. doi:10.11323/jjmp.36.1_35
27. Nicolae A, Morton G, Chung H, Loblaw A, Jain S, Mitchell D, et al. Evaluation of a machine-learning algorithm for treatment planning in prostate low-dose-rate brachytherapy. *Int J Radiat Oncol Biol Phys* (2017) 97:822–9. doi:10.1016/j.ijrobp.2016.11.036

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

SYSTEMATIC REVIEW
published: 14 May 2018
doi: 10.3389/fonc.2018.00131

Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review

*Amit Jethanandani[1,2], Timothy A. Lin[1,3], Stefania Volpe[1,4], Hesham Elhalawani[1], Abdallah S. R. Mohamed[1,5,6], Pei Yang[1,7] and Clifton D. Fuller[1,6]**

[1] Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States, [2] College of Medicine, The University of Tennessee Health Science Center, Memphis, TN, United States, [3] Baylor College of Medicine, Houston, TX, United States, [4] Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy, [5] Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, University of Alexandria, Alexandria, Egypt, [6] Graduate School of Biomedical Sciences, The University of Texas Health Science Center, Houston, TX, United States, [7] Hunan Cancer Hospital, Department of Head and Neck Radiation Oncology, Changsha, China

Background: Radiomics has been widely investigated for non-invasive acquisition of quantitative textural information from anatomic structures. While the vast majority of radiomic analysis is performed on images obtained from computed tomography, magnetic resonance imaging (MRI)-based radiomics has generated increased attention. In head and neck cancer (HNC), however, attempts to perform consistent investigations are sparse, and it is unclear whether the resulting textural features can be reproduced. To address this unmet need, we systematically reviewed the quality of existing MRI radiomics research in HNC.

Methods: Literature search was conducted in accordance with guidelines established by Preferred Reporting Items for Systematic Reviews and Meta-Analyses. Electronic databases were examined from January 1990 through November 2017 for common radiomic keywords. Eligible completed studies were then scored using a standardized checklist that we developed from Enhancing the Quality and Transparency of Health Research guidelines for reporting machine-learning predictive model specifications and results in biomedical research, defined by Luo et al. (1). Descriptive statistics of checklist scores were populated, and a subgroup analysis of methodology items alone was conducted in comparison to overall scores.

Results: Sixteen completed studies and four ongoing trials were selected for inclusion. Of the completed studies, the nasopharynx was the most common site of study (37.5%). MRI modalities varied with only four of the completed studies (25%) extracting radiomic features from a single sequence. Study sample sizes ranged between 13 and 118 patients (median of 40), and final radiomic signatures ranged from 2 to 279 features. Analyzed endpoints included either segmentation or histopathological classification parameters (44%) or prognostic and predictive biomarkers (56%). Liu et al. (2) addressed the highest number of our checklist items (total score: 48), and a subgroup analysis of methodology checklist items alone did not demonstrate any difference in scoring trends between studies [Spearman's $\rho = 0.94$ ($p < 0.0001$)].

OPEN ACCESS

Edited by:
Issam El Naqa, University of Michigan, United States

Reviewed by:
Marc van Hoof, Maastricht University Medical Centre (MUMC), Netherlands
Pavankumar Tandra, University of Nebraska Medical Center, United States

***Correspondence:**
Clifton D. Fuller
cdfuller@mdanderson.org

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *31 January 2018*
Accepted: *10 April 2018*
Published: *14 May 2018*

Citation:
Jethanandani A, Lin TA, Volpe S, Elhalawani H, Mohamed ASR, Yang P and Fuller CD (2018) Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review. Front. Oncol. 8:131. doi: 10.3389/fonc.2018.00131

Conclusion: Although MRI radiomic applications demonstrate predictive potential in analyzing diverse HNC outcomes, methodological variances preclude accurate and collective interpretation of data.

Keywords: radiomics, magnetic resonance imaging, MRI, texture analysis, head and neck, radiation oncology

INTRODUCTION

Rationale

Tumor characterization remains a major obstacle in the treatment of HNC patients (3, 4). Structural heterogeneity may represent underlying differences in tumor biology, which often cannot be explained by clinical data alone (5–8). Radiomics, the quantitative evaluation of anatomic structures from diagnostic imaging modalities, could possibly mitigate this variance (5, 6, 9). By describing morphological parameters and textural features from voxel elements, radiomics has the potential to examine tumors entirely (10–13).

Although multiple studies have applied radiomic analyses in HNC patients, computed tomography (CT) is the imaging modality most frequently investigated (14–26). This preference is due, in part, to the relative ease of data extraction and interpretation: Textural features can be derived from CT signal intensities (SIs) because their units of measurement, Hounsfield units (HUs), directly represent tissue radiodensity. Thus, SI gradients contain information about structural properties, which could then be translated into clinically meaningful data (9).

Computed tomography affords yet another advantage in that its imaging performance tends to be standardized across scanners and vendors (9). However, CT acquisition parameters can still influence the appearance of radiomic features (27). In non-small cell lung cancer (NSCLC), Mackin et al. (27) designed a radiomics-specific CT phantom to test inter-scanner variability. Mean CT number, reflected in HU, approximated the same variability between extracted tumor features from the scans themselves (27). Although extraction of features with discriminative ability from multiple scanners is promising, research is lacking in their application and robustness. Likewise, variances in reconstruction algorithms and image noise represent barriers to the accuracy of extracted features (9).

Similarly, radiomic studies based on magnetic resonance imaging (MRI) also face derivational challenges intrinsic to the technology. Not only are scanner parameters obstacles to reproducibility of features, but images themselves may reflect multiple tissue properties with specific acquisition characteristics (28). For instance, MRI SIs depends on pulse sequences, relaxation times, as well as a host of other acquisition-related processes; thus, seamless integration of radiomic analyses requires substantive effort (28).

When conducted appropriately, however, such studies can potentially provide a breadth of information superior to extrapolated values from CT radiomic features, as multiple physical properties of a voxel can be extracted *via* distinct sequence acquisition processes (e.g., spin–spin, proton density) and could be leveraged even further using novel techniques for simultaneous voxel characterization (e.g., MR fingerprinting) (29).

For example, MRI radiomics could potentially describe distinct patterns in tumor physiology: phenotypic categories from diffusion-weighted imaging (DWI) and dynamic contrast-enhanced (DCE) MRI have successfully predicted prognostic status in breast cancer patients (30). In addition, radiomic features derived from T1-weighted MRI reliably categorized molecular subtypes of breast tumors (31). For cases of glioblastoma (GBM), MRI radiomic profiles outperformed clinical and radiologic risk models in stratification of survival (32). Radiomic features have also successfully classified prostate tumors by Gleason scores (33, 34).

Objectives and Research Question

To the best of our knowledge, MRI radiomic applications in HNC have yet to be systematically summarized and reviewed in the clinical literature. In this effort, we assessed the quality of existing research: We comprehensively described MRI radiomic studies specific to the head and neck sub-site, with an intentional focus on study design. We compare and contrast the studies with a checklist based on Luo et al. (1) Enhancing the Quality and Transparency of Health Research (EQUATOR) methodology reporting guidelines. Subsequently, we discuss ongoing clinical trials and suggest future directions for MRI radiomic applications in HNC. The purpose of

Abbreviations: ADC, absolute diffusion coefficient; ARM, auto-regressive model; CCC, concordance correlation coefficient; ChiCTR, Chinese Clinical Trial Registry; CI, confidence interval; CT, computed tomography; DCE, dynamic contrast-enhanced; DICOM, digital imaging and communications in medicine; DWI, diffusion-weighted imaging; EQUATOR, Enhancing the Quality and Transparency of Health Research; FDG/PET, fludeoxyglucose-positron emission tomography; fMRI, functional magnetic resonance imaging; GBM, glioblastoma; GLAG, gray-level absolute gradient; GLCM, gray-level co-occurrence matrix; GLGCM, gray-level gradient co-occurrence matrix; GLH, gray-level histogram; GLRLM, gray-level run-length matrix; HNC, head and neck cancer; HU, Hounsfield unit; IBSI, image biomarker standardisation initiative; ICC, intraclass coefficient constant; IP, inverted papilloma; LAMBDA-[RAD]^2-HN initiative, a Large-scale Image Aggregation for Machine-Learning/Big Data Applications in Radiomics/Radiotherapy for Head and Neck Cancer; LDA, linear discriminant analysis; MDACC, MD Anderson Cancer Center; MRE, magnetic resonance elastography; MRI, magnetic resonance imaging; MS, methodology score; NCBI, National Center for Biotechnology Information; NIH RePORTER, National Institute of Health Research Portfolio Online Reporting Tool; NPC, nasopharyngeal cancer; NSCLC, non-small cell lung cancer; OPC, oropharyngeal cancer; PCA, principal component analysis; PFS, progression-free survival; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; QA, quality analysis; QIBA, Quantitative Imaging Biomarkers Alliance; QoL, quality of life; RECIST, Response Evaluation Criteria in Solid Tumors; ROI, region of interest; RT, radiotherapy; SCC, squamous cell carcinoma; SI, signal intensity; SNR, signal-to-noise ratio; STIR, short tau inversion recovery; SVM, support vector machine; TCIA, The Cancer Imaging Archive; TS, total score; WT, wavelet transform.

this systematic review is to assess the level of evidence and gauge the applicability of MRI radiomics in HNC.

METHODS

Study Design and Systematic Review Protocol

Study methodology followed outlines established by Preferred Reporting Items for Systematic Reviews and Meta-Analyses (**Figure 1**).

Eligibility Criteria

Full-text, original manuscripts, published in English, accepted for publication, and available online or in-print were evaluated. For inclusion, study populations consisted of patients diagnosed with HNC. All other cancer populations were excluded. Interventions included investigations of MRI radiomic features, where MRI was the primary imaging modality implemented. Studies exclusively researching first-order MRI features were excluded as they did not accurately represent the scope of typical MRI radiomic applications in HNC. Regarding outcomes, studies were included if they investigated segmentation accuracy, histopathological classification parameters, or prognostic and predictive biomarkers. Study design could be observational (e.g., prospective cohort, retrospective cohort, and case–control) or a clinical trial (e.g., randomized controlled trial).

Study Search Strategy and Process

Electronic databases (National Center for Biotechnology Information PubMed, Elsevier EMBASE, National Institute of

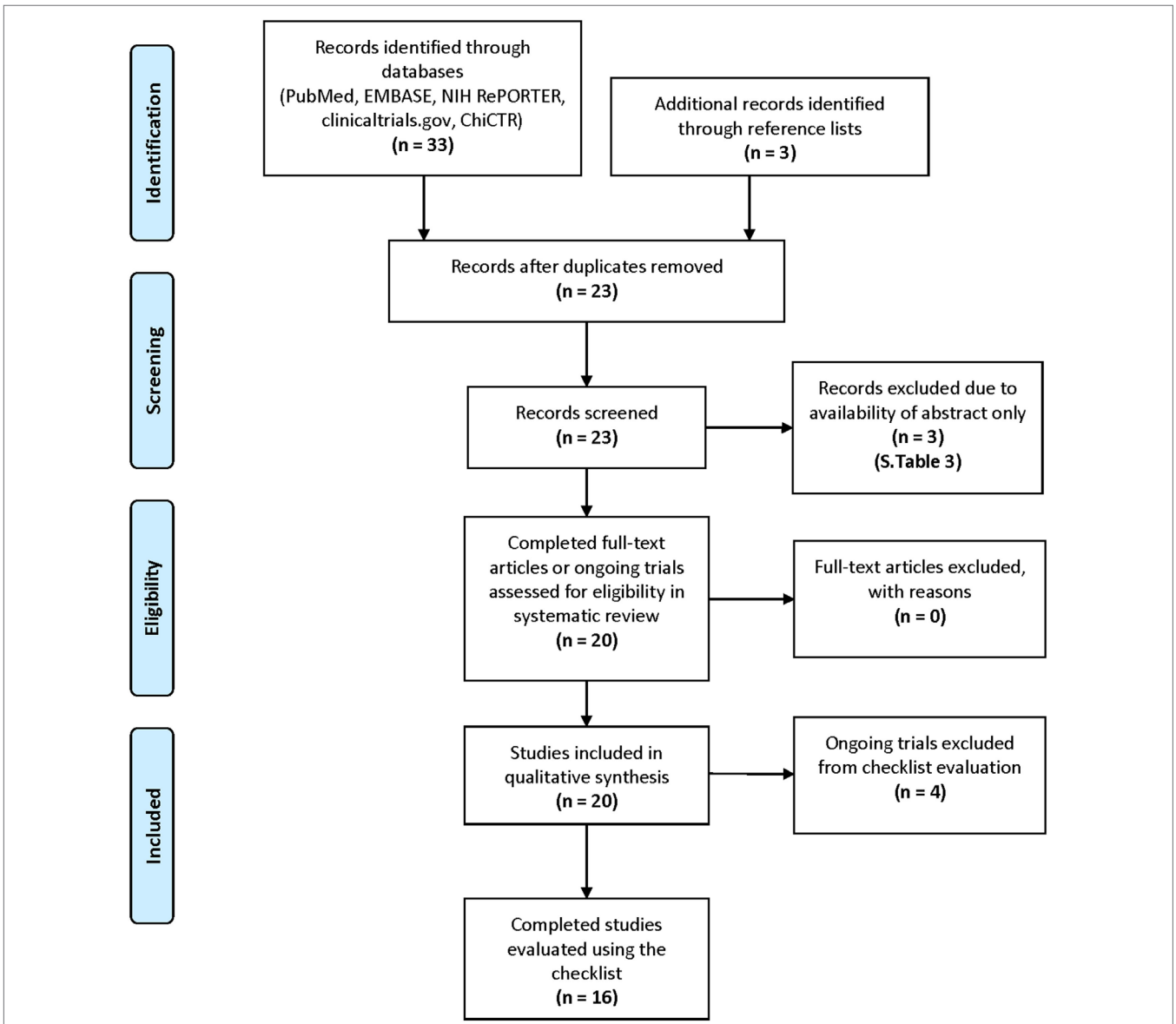

FIGURE 1 | Study methodology and search strategy *via* Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (35).

Health Research Portfolio Online Reporting Tool, ClinicalTrials.gov, and the Chinese Clinical Trial Registry) were searched from January 1990 through November 2017. Keywords and search strategy are described in our supplementary material (Table S5). For each included manuscript, reference lists were searched for additional eligible studies. Study search was completed by three authors independently (Amit Jethanandani, Timothy A. Lin, and Stefania Volpe), reviewing manuscripts in a stepwise method: By title alone, followed by abstract, then full-text. Search results were imported into individual spreadsheets using JMP Pro software version 12.1.0 (SAS Institute Inc., Cary, NC, USA). Discrepancies between results were discussed at team meetings, moderated by a fourth author (Hesham Elhalawani). Study search and selection were completed on November 13, 2017.

Data Sources, Study Sections, and Data Extraction

Selected studies consisted of completed research and ongoing trials. Once a final list was established, data extraction was completed independently by two authors (Amit Jethanandani and Timothy A. Lin) then assessed for quality by a third author (Hesham Elhalawani). Information was extracted into JMP Pro spreadsheets and included the following data: Manuscript title; authors; publication date; number of patients; head and neck sub-site; MRI modality and/or sequence used for radiomics analysis; region of interest (ROI) segmentation method; image pre-processing; feature extraction software; analyzed endpoint; statistical findings: radiomic model performance; conclusions; search terms and databases used to identify selected studies. Completed studies were stratified based on endpoints evaluated: Segmentation or histopathological classification vs. prognostic or predictive measures. Synthesis of data into a final spreadsheet was accomplished at team meetings among three authors (Amit Jethanandani, Timothy A. Lin, and Hesham Elhalawani).

Checklist Construction

A qualitative scoring method was developed for independent evaluation of completed studies. This system was adapted from Luo et al. (1) EQUATOR methodology reporting guidelines, which represent criteria outlined by a multidisciplinary panel of 11 clinicians, machine-learning specialists, and expert statisticians. The guidelines aimed to achieve two main objectives: (1) establish a list of key reporting items and (2) design a standardized, stepwise approach for generation of predictive models. The Delphi method was leveraged to iteratively narrow a list of included topics, discussed over e-mail between the panel members, to the final guidelines.

The guidelines were categorized by manuscript section for each reporting item: Title and abstract, introduction, methods, results, and discussion. Within these categories, reporting items were grouped by subsection. For example, the methods section contained the following groups: "Describe the setting," "define the prediction problem," "prepare data for model building," "build the predictive model," and "report the final model and performance." Our checklist mirrored this organization, with a few exceptions: Within the "build the predictive model" subsection, we further defined "data (feature) pre-processing" and "basic statistics of the dataset." Data pre-processing refers to data cleaning, data transformation, outlier removal, criteria for outlier removal, and handling of missing values. Basic statistics included items clarifying whether the model reflected the chosen classification or regression problem, the validation strategy, validation metrics, and the starting time for validation data collection. For organization of reporting items, a blank checklist is provided in our supplementary data section (Table S1 in Supplementary Material).

Each mandatory checklist item was categorized into a yes/no binary variable, which indicated whether the study appropriately addressed the corresponding criteria. The checklist was designed by one author (Timothy A. Lin) and subsequently revised by two authors (Amit Jethanandani and Hesham Elhalawani). Each completed study was scored individually by two authors (Amit Jethanandani and Timothy A. Lin). After all completed studies were scored, a group of three authors (Amit Jethanandani, Timothy A. Lin, and Hesham Elhalawani) met together to resolve discrepancies. There were 55 total checklist items, with two items containing sub-scores, representing a maximum overall score of 58 points. Once total checklist scores [total score (TS)] were finalized, methodology scores (MS) alone were generated for each completed study.

Data Analysis

Descriptive statistics for all included studies were populated and reviewed. For completed studies, TS and MS were tabulated in JMP Pro software. In addition, a subgroup analysis comparing collinearity of MS to TS was conducted using Spearman's ρ. Subgroup analysis was completed using the same JMP Pro software mentioned earlier.

RESULTS

Study Selection and Characteristics

Sixteen completed (2, 36–50) and four ongoing studies (51–54) were selected for inclusion. For completed studies, online or print publication dates ranged between May 2013 and October 2017. The selected studies could be retrieved from PubMed, and the most successful search term was "MRI texture analysis" (50% discovered with this keyword alone).

Synthesized Findings of Completed Studies

Patient sample sizes ranged between 13 and 118 patients with a median of 40 patients (**Table 1**). Head and neck sub-sites were diverse, including tumor volumes as well as normal anatomic structures. Of studies extracting radiomic features from tumor volumes, nasopharyngeal cancer (NPC) studies (37.5%) were the most common. Investigations of radiotherapy (RT)-related toxicities in normal tissue composed a small sample of the cohort (12.5%). Specific sub-sites were unknown for two studies (12.5%).

Magnetic resonance imaging sequences also varied, with T1-weighted, T2-weighted, and contrast-enhanced T1-weighted scans representing the most commonly used sequences. Only four studies (25%) derived texture features from a single MRI sequence. Thor et al. (45) extracted 24 textures, containing first- and second-order features, from T1-weighted post-contrast

images to quantify radiation-induced trismus. Brown et al. (36) investigated whether 21 texture features from a set of 300 DWI MRI parameters could reliably predict histopathological classification of thyroid tumors. Jansen et al. (40) generated pharmokinetic maps from DCE MRI images, applying texture measures of energy and homogeneity to determine associations with treatment response in oropharyngeal cancer patients.

Region of interest segmentation methods were less variable: Manual segmentation by trained experts alone (62.5%) composed the majority of studies. This was followed by combined manual and autosegmentation (31.25%), with one segmentation method unspecified (6.25%). One study investigated the classification performance of an autosegmentation method. Fruehwald-Pallamar et al. (38) leveraged a three-step strategy: Atlas-based registration, support vector machine (SVM) feature training, and parotid volume segmentation using trained feature SVM. For validation, reliability of the autosegmentation method was compared with trained physician contours using a Dice overlap ratio.

Most studies (62.5%) clarified image pre-processing steps before feature extraction. Preferred software for feature extraction included Matlab (37.5%) (MathWorks, Natick, MA, USA) and MaZda (25%) (Institute of Electronics, Technical University of Lodz, Poland). Feature pre-processing and model selection methods are discussed in the "Checklist scores" section of this manuscript.

Final radiomic signatures ranged from inclusion of 2 to 279 features. The upper limit reflects the choice of one study to maintain their initially derived feature set, which was not reduced in dimensionality. Meyer et al. (41) generated 279 features from T1-weighted and T2-weighted images corresponding to the following categories: gray-level co-occurrence matrix (GLCM), gray-level histogram, gray-level run-length matrix, gray-level absolute gradient, auto-regressive model, and wavelet transform. They then compared the derived T1- or T2-weighted features to cellular density, presence of Ki-67 antigen, or p53 index histopathology in 12 thyroid cancer patients.

Reports of radiomic model performance were typically positive (93.75%). However, Fruehwald-Pallamar et al. (39) concluded texture analysis was not practical across multiple MRI protocols, scanners, and vendors. **Table 1** lists the statistical findings specific to radiomic model performance of each study. Linear discriminant analysis (LDA) was the most commonly identified classification method, with four studies (25%) leveraging LDA to combine or reduce feature subsets. Likewise, four studies (25%) investigating progression outcomes in NPC patients utilized least absolute shrinking and Lasso methods to select significantly associated features for inclusion in final models. Only seven studies (44%) completely reported the predictive performance of their final model, in terms of their validation strategies, parameter estimates, and confidence intervals (CIs).

Analyzed endpoints ranged from segmentation and histopathological classification categories (44%) to prognostic or predictive biomarkers (56%). Among studies evaluating segmentation or classification, analyzed endpoints included: Histopathological classification (85.7%) and segmentation accuracy (14.3%). For studies assessing prognostic and predictive biomarkers, endpoints included: treatment response (33.3%), progression-free survival (PFS) (22.2%), progression dichotomized (22.2%), prognostic performance of predicting local or distant treatment failure (11.1%), and presence of radiation-induced trismus (11.1%).

All six NPC studies investigated prognostic or predictive biomarkers. Although they contained varying sample sizes (100–118), four studies (42, 47–49) selected from the same number of extracted radiomic features (970), subsequently constructing radiomic signatures from contrast-enhanced T1-weighted or T2-weighted feature categories. Among these studies, three investigated progression (either dichotomized yes/no or analyzed continuously) or a construct of prognostic performance. Liu et al. (2), alternatively investigated treatment response, defined using the Response Evaluation Criteria in Solid Tumors (RECIST). Patients with partial or complete response were considered responders, whereas patients with stable or progressive disease were classified as non-responders. One hundred and twenty six texture parameters were selected from contrast-enhanced T1-weighted, T1-weighted alone, and T2-weighted feature categories, then reduced to 15 features: GLCM, intensity size-zone matrix, and gray-level-gradient co-occurrence matrix. Using two separate selection methods, the remaining NPC study, Farhidzadeh et al. (50), examined the prognostic predictive power of intratumoral features—from either highly or weakly enhancing sub-regions—to classify patients by PFS category.

Checklist Scores

Finalized checklist scores are available in our supplementary dataset (Table S2 in Supplementary Material). Liu et al. (2) addressed the highest number of checklist items (TS: 48), followed by Brown et al. (36) and Ramkumar et al. (43) (TS: 45). Of note, all studies scored points for identifying their clinical goals, stating their predictive modeling, defining their target(s) of prediction, describing their sample size, defining the observational units of their response variable(s), interpreting their final model(s), and reporting the clinical implications of their data. By subsection, most study titles (93.75%) identified their reports as introducing a predictive model. Abstracts typically addressed objectives (87.5%), performance metrics in point estimates (87.5%), and practical relevance of study conclusions (87.5%); however, only three abstracts contained information on data sources (18.75%) or framed their performance metrics in terms of CIs (18.75%). Although only six study introductions addressed prediction accuracy of existing models (37.5%), this section contained the highest number of unanimously addressed items (50% of checklist items were unanimously addressed).

Methodology criteria contained the most checklist items [$n = 32$ (58.1%)]. Of the subsections in this category, studies missed the most points for failing to clarify their data (feature) pre-processing: Only seven studies (44%) discussed their data transformation, four (25%) removed outliers, three (18.75%) stated criteria for outlier removal, and one study (6.25%) discussed how missing values were handled. However, missing information in the abstract section, such as data sources, was eventually addressed in study methods (75%). Other common omissions included failures to specify model selection strategies (50% addressed); to define performance metrics in selecting the best model (37.5%); to explain the practical cost of prediction

TABLE 1 | Magnetic resonance imaging (MRI) radiomics in HNC: completed studies

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Studies on radiomics for segmentation and histopathological classification													
MRI texture analysis reflects histopathology parameters in thyroid cancer—a first preliminary study	Meyer HJ, Schob S, Hohn AK, Surov A	10/6/2017 (electronic publication, ePub); 12/2017 (Print)	13	Thyroid	T1-weighted turbo spin echo (TSE); T2-weighted TSE	Not specified	Yes	MaZda	Histopathological classification	279 texture features were analyzed for univariate association with histological parameters using a Spearman's correlation coefficient	Several significant correlations were identified between texture features and histopathology	2	1
Multi-institutional validation of a novel textural analysis tool for preoperative stratification of suspected thyroid tumors on diffusion-weighted MRI	Brown AM, Nagala S, McLean Ma, Lu Y, Scoffings D, Apte A, Gonen M, Stambuk HE, Shaha AR, Tuttle RM, Deasy JO, Priest AN, Jani P, Shukla-Dave A, Griffiths J	5/20/2015 (ePub); 4/2016 (Print)	42 (training=24, validation=18)	Thyroid	Diffusion-weighted imaging (DWI)	Manual	Yes	MaZda	Histopathological classification	A linear discriminant analysis (LDA) model of the top 21-ranking MaZda textural features classified 89/94 ROIs with 92% sensitivity and 96% specificity [AUC: 0.97, 95% confidence interval (CI): 0.92–1.0]. In a test set of 18 cases, the model's sensitivity was 89% (95% CI: 65–99%) and its specificity was 97% (95% CI: 74–100%)	Texture analysis is sensitive and specific for stratification of thyroid nodules	2	1
MRI texture analysis predicts p53 status in head and neck squamous cell carcinoma	Dang M, Lysack JT, Wu T, Matthews TW, Chandarana SP, Brockton NT, Bose P, Bansal G, Cheng H, Mitchell JR, Dort JC	9/25/2014 (ePub); 1/2015 (Print)	16	Oropharynx	Contrast-enhanced T1-weighted FSE; T2-weighted fast spin echo (FSE) with fat saturation; DWI	Manual	Yes	2D Fast Time-Frequency Transform Tool	Histopathological classification	A model of seven significant variables (determined using a subset-size forward selection algorithm and isolation of high-classification percentage variables) correctly classified 81.3% of tumors (κ: 0.625, $p < 0.05$)	A radiomic model containing variables with high classification performance could predict p53 status in oropharyngeal cancer patients	2	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Texture-based analysis of 100 MR examinations of head and neck tumors—is it possible to discriminate between benign and malignant masses in a multicenter trial?	Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME	9/30/2015 (ePub); 2/2016 (Print)	100	Head and neck benign (cysts = 8, inflammatory masses = 5, parotid = 9, glomus = 9, vascular malformation = 5, schwannoma = 4, other = 6) and malignant (squamous cell carcinoma = 31, lymphoma = 8, adenoid cystic = 5, adeno = 4, other = 6) tumors	Various	Manual and autosegmentation	No	MaZda	Histopathological classification	LDA models based off subsets of previously-identified, significant texture features demonstrated differences on STIR (61.29–80.65%) and T2-weighted images (T2-TSE: 81.82–100%, T2-TSE with fat suppresion: 71.74–78.26%) for 2D evaluation and on contrast-enhanced T1-TSE with fat saturation (58.54–85.37%) for 3D evaluation. Secondary analysis of subgroups by Tesla strength was also conducted	Texture analysis is not practical for differentiation of tumors using different magnetic resonance (MR) protocols on different MR scanners	2	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal MRI study in head-and-neck radiation therapy	Yang X, Wu N, Cheng G, Zhou Z, Yu DS, Beitler JJ, Curran WJ, Liu T	10/13/2014 (ePub); 12/2014 (Print)	15	Head and neck (oropharynx and larynx but other sites not specified)	Contrast-enhanced T1-weighted; Contrast-enhanced T2-weighted	Manual and autosegmentation	Yes	Not specified	Segmentation accuracy	A three-step autosegmentation method leveraging, as a component, a trained kernel-based support vector machine (SVM) model successfully differentiated 100% of parotid volumes where the average percentage of volume differences between the proposed method and manual physician contours were 7.98% (left parotid) and 8.12% (right parotid). Average Dice volume overlap: 91.1 ± 1.6% (left) and 90.5 ± 2.4% (right). Significant differences in volume reductions were found between 3-month and 1-year follow-up examinations ($p = 0.19$) and between 6-month and 1-year follow-up examinations ($p = 0.14$)	An autosegmentation method leveraging SVM models could accurately segment parotid glands when compared with manual review by trained experts	2	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Texture-based and diffusion-weighted discrimination of parotid gland lesions on MR images at 3.0 Tesla	Fruehwald-Pallamar J, Czerny C, Holzer-Fruehwald L, Nemec SF, Mueller-Mang C, Weber M, Mayerhoefer ME	5/23/2013 (ePub); 11/2013 (Print)	38	Parotid masses	Contrast-enhanced T1-weighted TSE; T1-weighted TSE; T1-weighted with fat suppression; Short Tau Inversion Recovery (STIR)	Manual and autosegmentation	Yes	MaZda	Histopathological classification	LDA models based off subsets of previously-identified, significant texture features was leveraged to determine differences between benign and malignant parotid masses or pleomorphic adenomas and Warthin tumors on multiple imaging modalities. Contrast-enhanced T1-weighted features correctly classified 81.8–84.5% of benign-malignant masses. Whereas, the same models applied to STIR imaging was poorer in distinguishing benign-malignant masses (73.5–78.4%) and pleomorphic adenomas-Warthin tumors (50–59%)	Contrast-enhanced T1-weighted features contained the most predictive textural information for distinguishing benign and malignant parotid masses. STIR images contained the least relevant textural information	2	1
MRI-based texture analysis to differentiate sinonasal squamous cell carcinoma from inverted papilloma	Ramkumar S, Ranjbar S, Ning S, Lal D, Zwart CM, Wood CP, Weindling SM, Wu T, Mitchell JR, Li J, Hoxworth JM	3/2/2017 (ePub); 5/2017 (Print)	46 (training=33, validation=13)	Sinonasal	Contrast-enhanced T1-weighted with fat suppression; T1-weighted; T2-weighted with fat suppression	Manual and autosegmentation	Yes	Python	Histopathological classification	The classification model, developed using five texture algorithms, demonstrated 90.9% accuracy in the training set and 84.6% accuracy in the validation set ($p = 0.537$). With both sets included, model accuracy (89.1%) outperformed neuroradiologists' ROI review (56.5%, $p = 0.0004$). This was not significantly different from neuroradiologist review of tumors (73.9%, $p = 0.060$) or entire images (87%, $p = 0.748$)	Machine-learning accuracy of texture analysis algorithms outperformed neuroradiologists' region of interest (ROI) review in classification of sinonasal carcinomas vs. inverted papillomas; however, its accuracy was not significantly different from neuroradiologists' review of tumors or entire images	2	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Studies on radiomics for prognostic and predictive biomarkers													
Exploration and validation of radiomics signature as an independent prognostic biomarker in stage III-IVb nasopharyngeal carcinoma	Ouyang FS, Guo B, Zhang B, Dong Y, Zhang L, Mo X, Huang W, Zhang S, Hu Q	9/26/2017 (ePub); 8/24/2017 (Print)	100 (training=70, validation=30)	Nasopharynx	Contrast-enhanced T1-weighted; T2-weighted	Manual	Yes	Matlab	PFS (Progression free survival)	In both the discovery and validation sets, a radiomic signature—using features selected via least absolute shrinkage and selection operator (Lasso) regression—successfully stratified patients by PFS risk category (HR: 5.14, $p < 0.001$; HR: 7.28, $p = 0.015$) while other identified clinical-pathologic risk factors for PFS were not significant (all p for HR > 0.05).	A radiomic signature based off pre-treatment MRI scans could predict PFS risk category and improve clinical decision-making	1	1
Advanced nasopharyngeal carcinoma: pre-treatment prediction of progression based on multi-parametric MRI radiomics	Zhang B, Ouyang FS, Gu D, Dong Y, Zhang L, Mo X, Huang W, Zhang S	9/22/2017 (ePub); 8/2/2017 (Print)	113 (training=80, validation=33)	Nasopharynx	Contrast-enhanced T1-weighted; T2-weighted	Manual	No	Matlab	Progression (Dichotomized to Yes and No categories)	Similar to the above strategy, radiomic features were selected using least absolute shrinkage and a Lasso method for significant association with progression. In both the training and validation cohort, the resulting radiomic-based model optimally performed when derived from combined contrast-enhanced T1-weighted and T2-weighted imaging (training: AUC: 0.896, 95% CI: 0.815–0.956; validation: 0.823, 95% CI: 0.645–1.00)	A radiomic model based on contrast-enhanced T1 and T2 features outperformed a model based on either MRI modality alone in its ability to predict progression in advanced nasopharyngeal cancer (NPC)	1	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma	Zhang B, He X, Ouyang FS, Gu D, Dong Y, Zhang L, Mo X, Huang W, Tian J, Zhang S	6/10/2017 (ePub); 9/10/2017 (Print)	110 (training=70, validation=40)	Nasopharynx	Contrast-enhanced T1-weighted; T2-weighted	Manual	Yes	Matlab	Prognostic performance of predicting local or distant treatment failure	Of the six feature selection and nine classification methods examined, the best predictive model utilized a combination Random Forest method (AUC: 0.8464 ± 0.0069; test error, 0.3135 ± 0.0088)	Radiomics models utilizing random forest methods demonstrated the highest prognostic performance compared with other machine-learning classification schemes, suggesting its utility in enhancing applications of radiomics in precision oncology	1	1
Radiomics features of multi-parametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma	Zhang B, Tian J, Dong D, Gu D, Dong Y, Zhang L, Lian Z, Liu J, Luo X, Pei S, Mo X, Huang W, Ouyang FS, Guo B, Liang L, Chen W, Liang C, Zhang S	3/9/2017 (ePub); 8/1/2017 (Print)	118 (training=88, validation=30)	Nasopharynx	Contrast-enhanced T1-weighted; T2-weighted	Manual	No	Matlab	PFS	Radiomic features were selected using least absolute shrinkage and a Lasso method for PFS nomograms. Radiomic signatures were significantly associated with PFS, with signatures derived from joint contrast-enhanced T1-weighted and T2-weighted images (Training C-index: 0.758, 95% CI: 0.661–0.856; Validation C-index: 0.737, 95% CI: 0.549–0.924). Outperforming signatures from either modality alone. When combined with clinical characteristics, the radiomics signature outperformed clinical characteristics alone in predicting PFS in advanced NPC (C-index, 0.776 vs. 0.649; $p < 1.60 \times 10^{-7}$)	Multiparametric MRI-based radiomic nomograms demonstrate prognostic ability in predicting progression in NPC patients	1	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Texture analysis on parametric maps derived from dynamic contrast-enhanced magnetic resonance imaging in head and neck cancer	Jansen JF, Lu Y, Gupta G, Lee NY, Stambuk HE, Mazaheri Y, Deasy JO, Shukla-Dave A	1/28/2016 (Print)	19	Oropharynx	Dynamic contrast-enhanced (DCE)	Manual	No	Matlab	Treatment response	Texture analysis on parametric DCE-MRI maps revealed energy of ve was higher in intra-treatment vs. pre-treatment scans ($p < 0.04$)	Pharmokinetic models performed on DCE images, producing ktrans and ve maps, were unable to predict treatment response. However, imaging biomarker E of ve was significantly higher in intra-treatment scans, vs. pre-treatment scans, suggesting a possible change in heterogeneity. The study ultimately conlcudes chemoradiation treatment reduces tumor heterogeneity in this patient cohort	2	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Use of texture analysis based on contrast-enhanced MRI to predict treatment response to chemoradiotherapy in nasopharyngeal carcinoma	Liu J, Mao Y, Li Z, Zhang D, Zhang Z, Hao S, Li B	1/18/2016 (ePub); 8/2016 (Print)	53 (training=42, validation=11)	Nasopharynx	Contrast-enhanced T1-weighted; T2-weighted; DWI; STIR TSE	Manual	Yes	Matlab	Treatment response	Three parameter sets of texture features derived from their respective imaging modalities were iteratively curated using multiple selection (e.g., the dynamic range metric) and classification methods (e.g., LDA). All three (T1: 0.952/0.939, T2: 0.904/0.905, DWI: 0.881/0.929) demonstrated an ability to predict treatment response, with supervised learning models using features from T1-weighted models exhibiting the highest classification performance vs. T2-weighted [artificial neural network (ANN): $p = 0.043$, k-nearest neighbors (k-NN): $p = 0.033$] or DWI (ANN: $p = 0.032$, k-NN: $p = 0.014$)	Radiomic models exhibit an ability to predict treatment response in NPC patients	2	1

(Continued)

TABLE 1 | Continued

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	Region of interest (ROI) segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance	Conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases [1 = PubMed, 2 = EMBASE, 3 =NIH, 4 = ClinicalTrials.Gov, 5 = Chinese Clinical Trial Registry (ChiCTR)]
Characterization of cervical lymph-nodes using a multi-parametric and multi-modal approach for an early prediction of tumor response to chemo-radiotherapy	Scalco E, Marzi S, Sanguineti G, Vidiri A, Rizzo G	9/14/2016 (ePub); 12/2016 (Print)	30	Head and neck (sites not specified)	T2-weighted; DWI; computed tomography (CT)	Manual	Yes	Python	Treatment response	Pre-treatment features outperformed mid-chemoradiation features in prediction of treatment response. Absolute diffusion coefficient (ADC) had the highest accuracy but, when combined with texture analysis, classification performance increased (accuracy = 82.8%). When T2-weighted texture features were evaluated independently, their best combination of pre-chemoradiation indices was equivalent in accuracy (81.8%)	An accurate assessment of response to chemoradiation in head and neck cancer patients could potentially be predicted from ADC parameters combined with texture analysis of T2-weighted imaging	2	1
Classification of progression free survival with nasopharyngeal carcinoma tumors	Farhidzadeh H, Kim JY, Scott JG, Goldgof DB, Hall LO, Harrison LB	3/24/2016 (ePub)	25	Nasopharynx	Contrast-enhanced T1-weighted	Manual and autosegmentation	No	Not specified	PFS (dichotomized)	Texture features derived from highly-enhancing signal intensity subregions classified PFS with 80% accuracy (AUC: 0.60). Texture features derived from weakly-enhancing subregions classified PFS with 76% accuracy (AUC: 0.76)	Intratumoral textural variations obtained through radiomics analyses can provide a "novel metric" to predict prognosis and assist clinicians in the design of individualized treatment regimens	1	1
A Magnetic Resonance Imaging-based approach to quantify radiation-induced normal tissue injuries applied to trismus in head and neck cancer	Thor M, Tyagi N, Hatzoglou V, Apte A, Saleh Z, Riaz N, Lee NY, Deasy JO	3/25/2017 (ePub); 1/2017 (Print)	20	Head and neck (sites not specified)	Contrast-enhanced T1-weighted	Manual	No	A Computational Environment for Radiotherapy Research	Radiation-induced trismus	Univariate statistical associations were derived. Mean dose to masseter (M), mean dose to medial pterygoid (MP), and Haralick correlation [gray-level co-occurrence matrix (GLCM)] of MP demonstrated the best discriminative ability in characterizing radiation-induced trismus (AUC: 0.85, 0.77, and 0.78, respectively)	An interplay between dose to M and MP as well as GLCM of MP suggests a possible relationship relevant to the etiology of radiation-induced trismus	1	1

errors (18.75%); and to identify which independent variables primarily take a single value (6.25%). Subgroup analysis of MS to TS demonstrated collinearity between both scoring sets [Spearman's $\rho = 0.94$ ($p < 0.0001$)].

Studies were strong in reporting their predictive performance, but only seven (44%) completely addressed their metrics in terms of validation strategies, parameter estimates, and CIs. A list of measured outcomes reported in each study is available in our supplementary material (Table S4). In addition, just one study (6.25%), Fruehwald-Pallamar et al. (38), compared their strategy with existing models in the literature using CIs. As for their conclusions, studies consistently failed to demonstrate whether sufficient data were available to fit their respective models (25%). However, most addressed potential bias (62.5%) as well as generalizability (68.75%) of their data.

Synthesized Findings of Ongoing Trials

Ongoing trials (51–54) (**Table 2**) estimate completion dates between June 2018 and December 2019 with one end-date unknown (25%). Three studies did not indicate a specific MRI sequence for feature extraction (75%). In addition, three studies will evaluate multiple head and neck sub-sites (75%). Two studies will prospectively evaluate data (50%), one study will be a case series (25%), and one study did not specify its design (25%). All studies will evaluate prognostic or predictive endpoints and, in addition, one study will evaluate a decision support system as its primary endpoint (25%). No preliminary data are available for any of the ongoing studies.

DISCUSSION

Summary of Main Findings

Our review represents the first attempt to summarize MRI radiomics research in HNC patients. Each completed study was evaluated using checklists generated from Luo et al. (1) EQUATOR methodology reporting guidelines: Individually scored, then collectively assessed for quality. Overall, our results indicate significant heterogeneity in study design, with limited consensus on a preferred radiomic signature. Thus, despite addressing reporting guidelines, included studies still demonstrate poor standardization. Such deficits may limit their generalizability and eventual use as clinical-decision support systems. However, this comprehensive review may improve comparison of data across study methodologies and structure similar analyses in other cancer sites.

Addressing Study Design

Several factors contribute to the lack of standardization across MRI radiomic studies in HNC patients. Variations follow the typical radiomics workflow: Patient populations (or head and neck sub-sites), image acquisition and pre-processing (MRI modalities), ROI segmentation methods, image pre-processing and feature extraction, feature selection, statistical modeling, and analyzed endpoints.

Head and Neck Sub-Sites

In our analysis, there was not a single head and neck sub-site representing a majority of all studies. However, the nasopharynx (37.5%) was the most commonly researched site. Diversity in head and neck sub-sites is not a unique characteristic of MRI radiomic studies, as research using CT radiomics has demonstrated a similar range of investigated patient populations (14). However, the high percentage of NPC studies may reflect the frequent use of MRI in their standard of care (55, 56).

In all six NPC studies, radiomic signatures demonstrated predictive potential. Of the feature categories included in their final radiomic signatures, GLCM was the only shared feature category between studies. This is consistent with NPC radiomic studies using other imaging modalities: Lu et al. (57) analyzed 88 texture features from FDG/PET-CT scans of 40 NPC patients, calculating the robustness of selected parameters in segmentation and discretization. Five GLCM properties (SumEntropy, Entropy, DifEntropy, Homogeneity1, and Homogeneity2) significantly demonstrated robustness at an intraclass coefficient constant $\geq$0.8 for seven segmentation methods and five discretization bin sizes.

Magnetic resonance imaging radiomics is not limited to studies of tumors alone. Radiomic signatures can predict RT-related toxicities in normal tissues, such as radiation-induced trismus (45), or they can be designed to autosegment parotid glands post-RT (46). Future studies should investigate whether radiomic features could predict the effects of RT-related toxicities on quality of life or if changes in corresponding critical organ volumes, such as structures involved in the swallowing mechanism, can be estimated.

MRI Modalities

Magnetic resonance imaging sequence preferences varied among studies, which is not uncommon to radiomics research in other cancer sites (58). Multiparametric approaches may reduce the risk of bias from features extracted from one sequence alone (49). However, since Brown et al. (36) and Jansen et al. (40) evaluated physiologic parameters, it is reasonable that additional MRI sequences would not adequately address their respective hypotheses. For example, Jansen et al. (40) selected DCE MRI for its ability to incorporate pharmacokinetic modeling. Before their study, DCE MRI parametric maps exhibited high image coherence among a tumor response group of limb sarcoma patients (59). Brown et al. (36) chose DWI MRI to improve its accuracy in stratification of thyroid nodules, a utility proven in feasibility studies (60, 61).

Other than sequence selection, MRI modalities may differ in their scanner properties, which would affect the reproducibility of images and, in turn, the texture features derived from them. To investigate whether texture-based signatures could appropriately classify head and neck masses across centers, Fruehwald-Pallamar et al. (39) recruited five MRI scanners from multiple manufacturers—each with varying field strengths, sequences, and acquisition parameters. The objective was to test whether texture analysis could be reliably reproduced in a "real world" clinical scenario. Although the authors ultimately could not recommend texture analysis for routine practice, certain texture features maintained discriminatory significance—particularly those derived from short tau inversion recovery and T2-weighted sequences. However, a review of study methodology revealed

TABLE 2 | Magnetic resonance imaging (MRI) radiomics in HNC: ongoing trials

Article title	Article authors	Publication date	Number of patients	Head and neck sub-site	MRI modality and/or sequence used for radiomics analysis	ROI segmentation method	Image pre-processing: yes/no	Feature extraction software	Analyzed endpoint	Statistical findings: radiomic model performance or conclusions	Successful search terms used [1 = Radiomic(s), 2 = MRI texture analysis, 3 = texture analysis, 4 = head and neck, 5 = magnetic resonance imaging texture analysis]	Databases (1 = PubMed, 2 = EMBASE, 3 = NIH, 4 = ClinicalTrials.Gov, 5 = ChiCTR)
Big data and models for personalized head and neck cancer decision support (BD2DECIDE)	Poli T, Schcekenback K, Schipper J, Colter L, Licitra L, Gatta G, Favales F, Trama A, De Cecco L, Silini EM, Maglietta G, Caminiti C, Iambin P, Hoebers F, Berlanga A	Estimated study completion date: 4/2019	Prospective arm: 450, Retrospective: 1000	Head and neck (Oral cavity, oropharynx, larynx, hypopharynx)	T1-weighted; T2-weighted; Computed Tomography (CT)	Not specified	Not specified	Not specified	Validation of decision support system; secondary outcomes include improved quality of life and assessment of survival time	N/A	1	4
Predictors of normal tissue response from the microenvironment in radiotherapy for prostate and head-and-neck cancer (MICROLEARNER)	Valdagni R, Orlandi E, Bedini N, Cecco LD, Zaffaroni N, Rancati T	Estimated study completion date: 12/31/2019	Prospective clinical trial population: 130 prostate, 130 HNC; prospective validation population: 70 prostate, 70 HNC	Prostate; Head and neck (oral cavity, pharynx, larynx, paranasal sinuses and nasal cavity, salivary glands)	MRI (not specified)	Not specified	Not specified	Not specified	Acute toxicity <90 days after Rt; secondary outcomes include late toxicity	N/A	1	4
Radiomics features for prediction of effect of local advanced nasopharyngeal carcinoma based on CT or MRI pre-chemoradiotherapy—a prospective cohort study	Su T-S	Estimated study completion date: TBD	Case series of 200	Nasopharynx	CT or MRI (not specified)	Not specified	Not specified	Not specified	Overall survival (OS), secondary outcomes include local-control rate and progression-free survival (PFS)	N/A	1	5
Personalized postoperative radiochemotherapy in patients with head and neck cancer	Zips DA	Estimated study completion date: 6/2018	Not specified	Head and neck (oropharynx and hypopharynx)	Positron Emission Tomography (PET), MRI (not specified)	Not specified	Not specified	Not specified	PFS; secondary outcomes—disease free survival, OS, development of a multi-parametric decision support system	N/A	1	4

omissions in model selection strategy, and their overall checklist score was below the median (TS: 37). Another issue was their intentionally diverse study population. Even though the sample consisted of 100 patients, the sub-sites were heterogeneous, with an unequal distribution of tumors among seven categories of benign masses and five categories of malignant masses. Thus, it is difficult to draw conclusions on radiomic signatures off this study alone.

Although the Quantitative Imaging Biomarkers Alliance (QIBA) continues to develop protocols for optimizing acquisition parameters, a technically confirmed profile for MRI radiomics does not exist. Yet, functional magnetic resonance imaging, DWI MRI, DCE MRI, and magnetic resonance elastography imaging biomarker profiles are currently in progress. The QIBA profile on DWI MRI (62), for example, specifies quality analysis (QA) of image acquisition and review of acquired data in brain, liver, and prostate studies. QIBA designed DWI MRI phantoms to streamline calculations of absolute diffusion coefficient (ADC) parametric maps and bias estimates, signal-to-noise ratios, as well as ADC spatial and *b*-value dependences. Extension of this protocol to DWI MRI radiomic studies in thyroid cancer could thus standardize ADC ROI assessment.

ROI Segmentation Methods

Once useable images are generated, ROIs must be segmented to assign volumes for feature derivation. Similar to other processes in the radiomics workflow, segmentation methods vary in their approach and design. Volumes are typically delineated either by manual contours, which can be laborious and time-consuming, or through autosegmenting machine-learning algorithms (63). Although the latter may present a new opportunity for standardized segmentation methods, challenges persist related to the complex anatomy of the head and neck sub-site, optimization of patient-based atlases, and SVM training characteristics (46). Further still, such methods may pale in comparison to recent advances in deep learning, where autosegmentation of myocardial volumes has already been accomplished on cardiac MRI (64). For studies leveraging one segmentation method alone, QA must be specified to limit ROI variation error. Example QA strategies include utilizing multiple experts to review volumes or statistically validating segmentation methods, as Fruehwald-Pallamar et al. (38) optimally demonstrated.

Image Pre-Processing and Feature Extraction

Before feature extraction, image quality should be ensured through pre-processing steps. To mitigate noise, which may confound raw imaging data, filters can be applied. Filter choice is dependent on acquisition parameters of imaging modalities, which necessitates standardization of preceding steps. Other obstacles to image pre-processing include diverse resampling schemes, varying computational definitions, motion artifacts, tumor size, and intratumoral heterogeneity, all of which need to be accounted for in study methodology (65, 66). As an example, Liu et al. (37) not only specified the standardization of their image acquisition parameters but also detailed their protocol for normalizing variations in image gray-level ranges.

Feature extraction ultimately depends on choice in software as well as characteristics of the features themselves. Radiomics features can be categorized by statistical output, where each subsequent ordinal group represents a higher complexity of voxel-based analysis. For example, first-order characteristics (e.g., ADC) are spatially independent descriptors of voxel distribution. Second-order characteristics, often equated with textural features, describe spatial relationships between two neighboring voxels (12). Often, however, studies do not explicitly characterize their extracted feature set, a major limitation to research reproducibility. At the minimum, the included studies in this review extracted spatially dependent features to investigate their endpoints.

Feature Selection

Each study developed a unique radiomic signature, which demonstrates both the strengths and weaknesses of "big data" research. Strengths include the volume of potentially useful quantitative information and flexibility of radiomic applications, but reproducibility and reliability of measured outcomes remain a concern (65). Thus, comparison of all selected features between studies is not entirely feasible. Although radiomic signatures contained similar categories of features, *diverse* parent feature samples derived from *diverse* MRI sequences with *their own diverse* scanner properties, signify the level of input and output variation inherent to these studies.

While most included studies detailed selection of extracted radiomic features, Meyer et al. (41) did not reduce their initially derived feature set. Direct and inverse correlations between specified features and classification parameters were discovered, but this presents a challenge to rationalize statistically. Potentially spurious associations (e.g., false positives) are inadequately addressed, which reflects the issues (e.g., approaches to data cleaning and transformation) identified collectively in our checklist. Future studies should clearly justify handling of missing values as well as terms and conditions for outlier removal. As checklist scores indicate, this remains an unaddressed issue.

Investigating the stability of MRI radiomic signatures could also identify necessary tweaks to the system. For instance, a feature selection method based on established stability criteria may help guide standardization of radiomic signatures (65). In soft tissue sarcomas, DWI MRI radiomic features derived from ADC maps were shown to maintain relevance across geometric transformations of ROIs (67). In recurrent GBM, test-retest reproducibility of 158 second-order radiomic features revealed 74% stability (68). Similarly, Liu et al. (2) only incorporated reproducible textural parameters in their final radiomic signature. They used a concordance correlation coefficient $\geq$0.9 to initially select features that maintained stability across different multi-observer ROI iterations of the same NPC patient. Outside of validation datasets, however, similar approaches are lacking in HNC studies.

Statistical Modeling

Discussed in previous reviews, a final radiomic signature is constrained by statistical analysis (9, 69, 70). When building predictive models, a set of candidate models should be reduced to the most

appropriate classifier, defined by performance metrics of a specific selection strategy (e.g., *k*-fold validation) (1, 66). Otherwise, a concern may be the adoption of dimensionality-reduction techniques solely to limit over-fitting of data. A combined feature extraction and statistical learning platform, built for radiomic challenges, would quell concerns about optimization of radiomic models. Until then, the aforementioned barriers persist across imaging modalities, with limited research focused exclusively on MRI radiomic applications (65).

Analyzed Endpoints

Choice of analyzed endpoint guides investigators through their specific radiomics pipeline. Thus, this adds another layer of complexity to selection, extraction, and modeling of features. To objectively predict outcomes, then, automating the above steps may preclude confounded associations. In their prospective MRI radiomic analysis of head and neck tumor p53 classification, for example, Dang et al. (37) used separate software for feature quantification and selection to identify best candidate predictors. Textural features can be biased by imbalances in events or classification parameters, particularly for prediction of rare outcomes. Statistical sampling techniques to enhance prediction accuracy should be implemented for unbalanced datasets.

In their 2016 review of HNC radiomics, Wong et al. (14) identified four of the included studies in our cohort, with three (75%) investigating classification schemes and just one (25%) analyzing prognostic or predictive biomarkers. At the time, CT radiomics research in HNC concentrated on the latter category (14). Discovered through our search strategy, abstracts from conference proceedings (Table S3 in Supplementary Material) all focused on prognostic endpoints in NPC patients (71–73). Thus, perhaps, MRI radiomic studies in HNC are trending toward these outcome measures.

Checklist Scores

Studies with the highest overall scores [e.g., Liu et al. (37) (TS: 48)] addressed more of the methodology reporting guidelines than studies with lower scores (Spearman's $\rho = 0.94$), which reflects areas of improvement for subsequent work. For example, Liu et al. (2) (MS: 30), were awarded points across the category except for one item (stating how missing values were handled). In addition to an internal 10-fold cross-validation strategy, the study externally validated their findings in an independent sample of 11 patients. They were also the only study to address each item in the "Build the predictive model" subsection. Their manuscript's discussion received points for every item in the "limitations" subsection; in particular, the authors demonstrated sufficient data available for fitting of their models (neglected in 75% of studies).

Likewise, Ramkumar et al. (43) addressed methodology items commonly missing in other studies. For instance, the authors explained possible prediction errors of texture analysis in distinguishing sinonasal squamous cell carcinoma from inverted papilloma. Similarly, they addressed multiple items in the data pre-processing subsection including data cleaning (e.g., feature reduction) and data transformation. The study meticulously described organization and selection of features, *via* a principal component analysis, as well as the metrics in building their final model. Although not technically an external validation set, the addition of a neuroradiologist review to an internal leave-one-out cross-validation assess buffered the strength of their classification accuracy.

Limitations

The review does present some notable limitations. A literature search with a known end-date may miss studies published in the interim; this is a limitation of any systematic review. Since MRI radiomics is a field still in its infancy, with a nomenclature not fully standardized, search keywords based on existing literature may not detect all eligible works inclusively. Specifically, keywords containing "texture analysis" may not encompass the breadth of radiomic investigations. To address this, we combed references of each included manuscript. Yet, we are aware of the challenges and risk of bias in selecting potential studies for inclusion and presenting a complete summary of a burgeoning research topic.

Although our checklist was constructed from established guidelines (1), the scoring system required multiple revisions to fairly assess the included studies. As the guidelines were not intended to be quantitative measurements, our group met frequently to weight each item. In addition, we removed guidelines which were difficult to interpret among all authors. Finally, we cannot predict whether the original authors of the guidelines would have constructed the same checklist. We can, however, attest to its quality, given its review by multiple expert radiation oncologists trained in radiomic analyses.

Conclusion

Magnetic resonance imaging radiomic studies in HNC lack standardization of study design, which practically limits their clinical relevance. Nonetheless, radiomic applications have demonstrated predictive potential in classification schemes and prognostic biomarker identification. Our quantitative scoring system may encourage routine study assessment, perhaps ensuring better data moving forward.

As our collation of the available HNC evidence indicates, MRI radiomics is an evolving field of study. Thus, we suggest several steps for streamlining future investigations. At our institution, novel radiomic-specific MRI phantoms are currently in development and may quantify the effects of inter-scanner variability on radiomic feature generation (70). Understanding the interplay between these processes will hopefully enhance data output. Regarding extraction and selection of features, the imaging biomarker standardisation initiative continues to derive testable categories (74). However, feature stability assessments in MRI are still pending. Analysis should be conducted using readily available software with sufficient flexibility across statistical platforms. Reports of finalized results should follow Luo et al. (1) EQUATOR methodology reporting guidelines.

To cross-validate radiomic signatures externally, tests should be performed on public patient datasets (e.g., The Cancer Imaging Archive). To this end, an upcoming multi-site collaboration between MDACC and other academic cancer centers will generate a repository of patient data in Digital Imaging and Communications in Medicine format, as part of our LAMBDA-[RAD]^2-HN initiative: a Large-scale Image Aggregation for Machine-Learning/Big Data

Applications in Radiomics/Radiotherapy for Head and Neck Cancer. This working group aims to provide an open-access library of curated "big data," rigorously maintained and routinely assessed for quality (75). Therefore, subsequent efforts to standardize MRI radiomics in HNC would share a reliable data pool.

AUTHOR CONTRIBUTIONS

Study designed by all authors. Literature search performed by AJ, TL, and SV. Data extraction completed by AJ and TL. Quality check completed by HE. Data synthesis of selected studies completed by AJ, TL, and HE. All tables formatted by AJ. Checklist designed by TL. Checklist structure revised by AJ and HE. Checklist scores for each study calculated by AJ and TL. Discrepancies between author checklist scores resolved by AJ, TL, and HE. Consort diagram designed by TL. Abstract drafted by SV, HE, and AJ. Cover letter and manuscript drafted by AJ. Abstract, cover letter, and manuscript reviewed and edited by SV, TL, HE, AM, PY, and CF.

FUNDING

CF: this research is supported by the Andrew Sabin Family Foundation; CF is a Sabin Family Foundation Fellow. CF receives funding and salary support from NIH, including: the National Institute for Dental and Craniofacial Research Award (1R01DE025248-01/R56DE025248-01); AM also receives funding from the National Institute for Dental and Craniofacial Research Award. a National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); the NIH Big Data to Knowledge (BD2K) Program of the National Cancer Institute (NCI) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825-01); NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148-01); an NIH/NCI Cancer Center Support Grant (CCSG) Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672); and an NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50 CA097007-10). CF has received direct industry grant support and travel funding from Elekta AB. HE is supported in part by the philanthropic donations from the Family of Paul W. Beach to Dr. G. Brandon Gunn. AJ, Dunagan Scholar, is supported by the Dunagan MD Medical Education Fund through The University of Tennessee Health Science Center, College of Medicine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fonc.2018.00131/full#supplementary-material.

TABLE S1 | Blank checklist.

TABLE S2 | Finalized checklist scores.

TABLE S3 | MRI radiomics in HNC: abstracts only.

TABLE S4 | Reports of measured outcomes.

TABLE S5 | Search strategy.

REFERENCES

1. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* (2016) 18(12):e323. doi:10.2196/jmir.5870
2. Liu J, Mao Y, Li Z, Zhang D, Zhang Z, Hao S, et al. Use of texture analysis based on contrast-enhanced MRI to predict treatment response to chemoradiotherapy in nasopharyngeal carcinoma. *J Magn Reson Imaging* (2016) 44(2):445–55. doi:10.1002/jmri.25156
3. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science* (2011) 333(6046):1157–60. doi:10.1126/science.1208130
4. The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* (2015) 517:576. doi:10.1038/nature14129
5. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* (2014) 5:4006. doi:10.1038/ncomms5006
6. Davnall F, Yip CSP, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging* (2012) 3(6):573–89. doi:10.1007/s13244-012-0196-6
7. O'Connor JPB, Rose CJ, Waterton JC, Carano RAD, Parker GJM, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res* (2015) 21(2):249–57. doi:10.1158/1078-0432.CCR-14-0990
8. Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. Computed tomography radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma. *Int J Radiat Oncol Biol Phys* (2017) 99(4):921–8. doi:10.1016/j.ijrobp.2017.06.002
9. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. QIN "Radiomics: the process and the challenges". *Magn Reson Imaging* (2012) 30(9):1234–48. doi:10.1016/j.mri.2012.06.010
10. Parmar C, Leijenaar RTH, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci Rep* (2015) 5:11044. doi:10.1038/srep11044
11. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. *Tomography* (2016) 2(4):430–7. doi:10.18383/j.tom.2016.00235
12. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* (2015) 278(2):563–77. doi:10.1148/radiol.2015151169
13. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48(4):441–6. doi:10.1016/j.ejca.2011.11.036
14. Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. *Transl Cancer Res* (2016) 5(4):371–82. doi:10.21037/tcr.2016.07.18
15. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to human papillomavirus status. *Oral Oncol* (2017) 71:150–5. doi:10.1016/j.oraloncology.2017.06.015
16. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar R, et al. Predictive and prognostic value of CT based radiomics signature in head and neck squamous cell carcinoma patients treated with concurrent chemoradiation therapy or bioradiation therapy and its added value to human papillomavirus

status. *Int J Radiat Oncol Biol Phys* (2017) 99(2):S13. doi:10.1016/j.ijrobp.2017.06.047
17. Fujita A, Buch K, Li B, Kawashima Y, Qureshi MM, Sakai O. Difference between HPV-positive and HPV-negative non-oropharyngeal head and neck cancer: texture analysis features on CT. *J Comput Assist Tomogr* (2016) 40(1):43–7. doi:10.1097/RCT.0000000000000320
18. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol* (2015) 5:272. doi:10.3389/fonc.2015.00272
19. Leijenaar RTH, Carvalho S, Hoebers FJP, Aerts HJWL, van Elmpt WJC, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol* (2015) 54(9):1423–9. doi:10.3109/0284186X.2015.1061214
20. Buch K, Fujita A, Li B, Kawashima Y, Qureshi MM, Sakai O. Using texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinomas on CT. *Am J Neuroradiol* (2015) 36(7):1343–8. doi:10.3174/ajnr.A4285
21. Zhang H, Graham CM, Elci O, Griswold ME, Zhang X, Khan MA, et al. Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. *Radiology* (2013) 269(3):801–9. doi:10.1148/radiol.13130110
22. Scalco E, Fiorino C, Cattaneo GM, Sanguineti G, Rizzo G. Texture analysis for the assessment of structural changes in parotid glands induced by radiotherapy. *Radiother Oncol* (2013) 109(3):384–7. doi:10.1016/j.radonc.2013.09.019
23. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol* (2013) 52(7):1391–7. doi:10.3109/0284186X.2013.812798
24. Raja J, Khan M, Ramachandra V, Al-Kadi O. Texture analysis of CT images in the characterization of oral cancers involving buccal mucosa. *Dentomaxillofac Radiol* (2012) 41(6):475–80. doi:10.1259/dmfr/83345935
25. Yu H, Caldwell C, Mah K, Poon I, Balogh J, MacKenzie R, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT Images. *Int J Radiat Oncol Biol Phys* (2009) 75(2):618–25. doi:10.1016/j.ijrobp.2009.04.043
26. Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. *IEEE Trans Med Imaging* (2009) 28(3):374–83. doi:10.1109/TMI.2008.2004425
27. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol* (2015) 50(11):757–65. doi:10.1097/RLI.0000000000000180
28. Zhao B, Tan Y, Tsai W-Y, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep* (2016) 6:23428. doi:10.1038/srep23428
29. European Society of Radiology (ESR). Magnetic Resonance Fingerprinting – a promising new approach to obtain standardized imaging biomarkers from MRI. *Insights Imaging* (2015) 6(2):163–5. doi:10.1007/s13244-015-0403-3
30. Maforo N, Li H, Lan L, Edwards A, Giger ML. SU-F-R-26: prognostic radiomics of breast cancer on DCE and DWI MR images. *Med Phys* (2016) 43(6Part6):3378. doi:10.1118/1.4955798
31. Li H, Zhu Y, Burnside ES, Huang E, Drukker K, Hoadley KA, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *NPJ Breast Cancer* (2016) 2:16012. doi:10.1038/npjbcancer.2016.12
32. Kickingereder P, Burth S, Wick A, Götz M, Eidel O, Schlemmer H-P, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology* (2016) 280(3):880–9. doi:10.1148/radiol.2016160845
33. Larue RTHM, Defraene G, Ruysscher DD, Lambin P, Elmpt WV. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol* (2017) 90(1070):20160665. doi:10.1259/bjr.20160665
34. Gnep K, Fargeas A, Gutiérrez-Carvajal RE, Commandeur F, Mathieu R, Ospina JD, et al. Haralick textural features on T2-weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. *J Magn Reson Imaging* (2017) 45(1):103–17. doi:10.1002/jmri.25335
35. Moher D, Liberati A, Tetzlaff J, Altman DG; The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* (2009) 6(7):e1000097. doi:10.1371/journal.pmed.1000097
36. Brown AM, Nagala S, McLean MA, Lu Y, Scoffings D, Apte A, et al. Multi-institutional validation of a novel textural analysis tool for preoperative stratification of suspected thyroid tumors on diffusion-weighted MRI. *Magn Reson Med* (2016) 75(4):1708–16. doi:10.1002/mrm.25743
37. Dang M, Lysack JT, Wu T, Matthews TW, Chandarana SP, Brockton NT, et al. MRI texture analysis predicts p53 status in head and neck squamous cell carcinoma. *AJNR Am J Neuroradiol* (2015) 36(1):166–70. doi:10.3174/ajnr.A4110
38. Fruehwald-Pallamar J, Czerny C, Holzer-Fruehwald L, Nemec SF, Mueller-Mang C, Weber M, et al. Texture-based and diffusion-weighted discrimination of parotid gland lesions on MR images at 3.0 Tesla. *NMR Biomed* (2013) 26(11):1372–9. doi:10.1002/nbm.2962
39. Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME. Texture-based analysis of 100 MR examinations of head and neck tumors – is it possible to discriminate between benign and malignant masses in a multicenter trial? *Fortschr Röntgenstr* (2016) 188(02):195–202. doi:10.1055/s-0041-106066
40. Jansen JFA, Lu Y, Gupta G, Lee NY, Stambuk HE, Mazaheri Y, et al. Texture analysis on parametric maps derived from dynamic contrast-enhanced magnetic resonance imaging in head and neck cancer. *World J Radiol* (2016) 8(1):90–7. doi:10.4329/wjr.v8.i1.90
41. Meyer H-J, Schob S, Höhn AK, Surov A. MRI texture analysis reflects histopathology parameters in thyroid cancer – a first preliminary study. *Transl Oncol* (2017) 10(6):911–6. doi:10.1016/j.tranon.2017.09.003
42. Ouyang F-S, Guo B-L, Zhang B, Dong Y-H, Zhang L, Mo X-K, et al. Exploration and validation of radiomics signature as an independent prognostic biomarker in stage III-IVb nasopharyngeal carcinoma. *Oncotarget* (2017) 8(43):74869–79. doi:10.18632/oncotarget.20423
43. Ramkumar S, Ranjbar S, Ning S, Lal D, Zwart CM, Wood CP, et al. MRI-based texture analysis to differentiate sinonasal squamous cell carcinoma from inverted papilloma. *AJNR Am J Neuroradiol* (2017) 38(5):1019–25. doi:10.3174/ajnr.A5106
44. Scalco E, Marzi S, Sanguineti G, Vidiri A, Rizzo G. Characterization of cervical lymph-nodes using a multi-parametric and multi-modal approach for an early prediction of tumor response to chemo-radiotherapy. *Phys Med* (2016) 32(12):1672–80. doi:10.1016/j.ejmp.2016.09.003
45. Thor M, Tyagi N, Hatzoglou V, Apte A, Saleh Z, Riaz N, et al. A magnetic resonance imaging-based approach to quantify radiation-induced normal tissue injuries applied to trismus in head and neck cancer. *Phys Imaging Radiat Oncol* (2017) 1:34–40. doi:10.1016/j.phro.2017.02.006
46. Yang X, Wu N, Cheng G, Zhou Z, Yu DS, Beitler JJ, et al. Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal MRI study in head-and-neck radiation therapy. *Int J Radiat Oncol Biol Phys* (2014) 90(5):1225–33. doi:10.1016/j.ijrobp.2014.08.350
47. Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, et al. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett* (2017) 403:21–7. doi:10.1016/j.canlet.2017.06.004
48. Zhang B, Ouyang F, Gu D, Dong Y, Zhang L, Mo X, et al. Advanced nasopharyngeal carcinoma: pre-treatment prediction of progression based on multi-parametric MRI radiomics. *Oncotarget* (2017) 8(42):72457–65. doi:10.18632/oncotarget.19799
49. Zhang B, Tian J, Dong D, Gu D, Dong Y, Zhang L, et al. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clin Cancer Res* (2017) 23(15):4259–69. doi:10.1158/1078-0432.CCR-16-2910
50. Farhidzadeh H, Kim JY, Scott JG, Goldgof DB, Hall LO, Harrison LB, editors. Classification of progression free survival with nasopharyngeal carcinoma tumors. *Conference proceedings: SPIE Medical Imaging*. SPIE (2016).
51. ClinicalTrials.gov [Internet]. *Identifier NCT02832102. Big Data and Models for Personalized Head and Neck Cancer Decision Support (BD2DECIDE)*. Bethesda, MD: National Library of Medicine (US) (2000). [cited 2017 Jan 2]. Available from: https://clinicaltrials.gov/ct2/show/NCT02832102 (Accessed: July 14, 2016).

52. ClinicalTrials.gov [Internet]. *Identifier NCT03294122. Predictors of Normal Tissue Response From the Microenvironment in Radiotherapy for Prostate and Head-and-Neck Cancer (MICROLEARNER)*. Bethesda, MD: National Library of Medicine (US) (2000). [cited 2017 Jan 2]. Available from: https://clinicaltrials.gov/ct2/show/NCT03294122 (Accessed: October 3, 2017).
53. Chinese Clinical Trial Register [Internet]. *Identifier ChiCTR-POC-17012506. Radiomics Features for Prediction of Effect of Local Advanced Nasopharyngeal Carcinoma Based on CT or MRI Pre-Chemoradiotherapy-A Prospective Cohort Study*. Chengdu, Sichuan: Ministry of Health (China) (2007). [cited 2017 Jan 2]. Available from: http://www.chictr.org.cn/showprojen.aspx?proj=21369 (Accessed: August 31, 2017).
54. ClinicalTrials.gov [Internet]. *Identifier NCT02666885. Personalised Postoperative Radiochemotherapy in Patients With Head and Neck Cancer*. Bethesda, MD: National Library of Medicine (US) (2000). [cited 2017 Jan 2]. Available from: https://clinicaltrials.gov/ct2/show/NCT02666885 (Accessed: January 28, 2016).
55. Sung SY, Kang MK, Kay CS, Keum KC, Kim SH, Kim Y-S, et al. Patterns of care for patients with nasopharyngeal carcinoma (KROG 11-06) in South Korea. *Radiat Oncol J* (2015) 33(3):188–97. doi:10.3857/roj.2015.33.3.188
56. King AD, Vlantis AC, Bhatia KSS, Zee BCY, Woo JKS, Tse GMK, et al. Primary Nasopharyngeal carcinoma: diagnostic accuracy of MR imaging versus that of endoscopy and endoscopic biopsy. *Radiology* (2011) 258(2):531–7. doi:10.1148/radiol.10101241
57. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of radiomic features in [11C]choline and [18F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. *Mol Imaging Biol* (2016) 18(6):935–45. doi:10.1007/s11307-016-0973-6
58. Stoyanova R, Takhar M, Tschudi Y, Ford JC, Solórzano G, Erho N, et al. Prostate cancer radiomics and the promise of radiogenomics. *Transl Cancer Res* (2016) 5(4):432–47. doi:10.21037/tcr.2016.06.20
59. Alic L, van Vliet M, van Dijke CF, Eggermont AM, Veenland JF, Niessen WJ. Heterogeneity in DCE-MRI parametric maps: a biomarker for treatment response? *Phys Med Biol* (2011) 56(6):1601–16. doi:10.1088/0031-9155/56/6/006
60. Shi HF, Feng Q, Qiang JW, Li RK, Wang L, Yu JP. Utility of diffusion-weighted imaging in differentiating malignant from benign thyroid nodules with magnetic resonance imaging and pathologic correlation. *J Comput Assist Tomogr* (2013) 37(4):505–10. doi:10.1097/RCT.0b013e31828d28f0
61. Chen L, Xu J, Bao J, Huang X, Hu X, Xia Y, et al. Diffusion-weighted MRI in differentiating malignant from benign thyroid nodules: a meta-analysis. *BMJ Open* (2016) 6(1):e008413. doi:10.1136/bmjopen-2015-008413
62. Diffusion-Weighted Imaging Task Force subgroup of the Perfusion Diffusion and Flow (PDF) Biomarker Committee. *QIBA Profile: Diffusion-Weighted Magnetic Resonance Imaging (DWI), Quantitative Imaging Biomarkers Alliance. Version 1.45. Profile Stage: Comment Resolution*. QIBA (2017). Available from: http://qibawiki.rsna.org/images/1/1d/QIBADWIProfilev1.45_20170427_v5_accepted.pdf (Accessed: January 10, 2018).
63. Heye T, Merkle EM, Reiner CS, Davenport MS, Horvath JJ, Feuerlein S, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part II. comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. *Radiology* (2013) 266(3):812–21. doi:10.1148/radiol.12120255
64. Curiale A, Colavecchia F, Kaluza P, Isoardi R, Mato G. Automatic myocardial segmentation by using a deep learning network in cardiac MRI. *2017 XLII Latin American Computer Conference (CLEI)*; 2017 Sept 4–8; Cordoba, Argentina. IEEE (2017). doi:10.1109/CLEI.2017.8226420
65. Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol* (2016) 61(13):R150–66. doi:10.1088/0031-9155/61/13/R150
66. Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuze S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol* (2017) 1(6):1191–206. doi:10.1093/annonc/mdx034
67. Bologna M, Montin E, Corino VDA, Mainardi LT. Stability assessment of first order statistics features computed on ADC maps in soft-tissue sarcoma. *Conf Proc IEEE Eng Med Biol Soc* (2017) 2017:612–5. doi:10.1109/EMBC.2017.8036899
68. Shiri I, Abdollahi H, Shaysteh S, Mahdavi S. Test-Retest Reproducibility and Robustness Analysis of Recurrent Glioblastoma MRI Radiomics Texture Features. Iranian Journal of Radiology (2017) (5):e48035. doi:10.5812/iranjradiol.48035
69. Ranjbar S, Ross Mitchell J. *Chapter 8 – An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine. Biomedical Texture Analysis*. Academic Press (2017). p. 223–45.
70. Yang J, Steinmann A, Mackin D, Stafford R, Followill D, Li J, et al. TU-H-FS4-9: development of An MRI Radiomics Phantom. *Med Phys* (2017) 44(6):6.
71. Nair JKR, Vallieres M, Shenouda G, Zeitouni A, Chankowsky J. Radiomics model from volumetric MRI high order texture analysis for pre-treatment stratification of patients with nasopharyngeal carcinoma. *Conference Proceedings: American Society of Head and Neck Radiology*. ASHNR (2016).
72. Zhang B. Multi-parametric MRI radiomics for pre-treatment prediction of the progression-free survival in advanced nasopharyngeal carcinoma. *Conference Proceedings: International Society for Magnetic Resonance in Medicine*. ISMRM (2017).
73. Ming X, Ying H, Huang R, Wang J, Hu W, Zhang Z, et al. MRI based radiomics signature, a quantitative prognostic biomarker for nasopharyngeal carcinoma. *Conference Proceedings: American Association of Physicists in Medicine*. AAPM (2017).
74. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative (2016). eprint arXiv:1612.07003.
75. Elhalawani H, Elgohari B, Yang P, Mohamed A, Zhang X, Fuller CD. A Cloud-based Platform for Large Scale Image Aggregation for Machine-learning/Big Data Applications in Radiomics/Radiotherapy for Head and Neck Cancer (LAMBDA-RAD2): Towards FAIR Data Sharing. Accepted for poster presentation at: AMIA 2018 Clinical Informatics Conference (May 9 2018). https://cic2018.zerista.com/event/member/474684

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 14 June 2018
doi: 10.3389/fonc.2018.00215

Deep Learning Renal Segmentation for Fully Automated Radiation Dose Estimation in Unsealed Source Therapy

Price Jackson[1,2,3], Nicholas Hardcastle[3], Noel Dawe[4], Tomas Kron[3], Michael S. Hofman[2] and Rodney J. Hicks[2]*

[1]*Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC, Australia,* [2]*Department of Molecular Imaging, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia,* [3]*Department of Physical Sciences, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia,* [4]*School of Physics, University of Melbourne, Melbourne, VIC, Australia*

Background: Convolutional neural networks (CNNs) have been shown to be powerful tools to assist with object detection and—like a human observer—may be trained based on a relatively small cohort of reference subjects. Rapid, accurate organ recognition in medical imaging permits a variety of new quantitative diagnostic techniques. In the case of therapy with targeted radionuclides, it may permit comprehensive radiation dose analysis in a manner that would often be prohibitively time-consuming using conventional methods.

Methods: An automated image segmentation tool was developed based on three-dimensional CNNs to detect right and left kidney contours on non-contrast CT images. Model was trained based on 89 manually contoured cases and tested on a cohort of patients receiving therapy with ^{177}Lu-prostate-specific membrane antigen-617 for metastatic prostate cancer. Automatically generated contours were compared with those drawn by an expert and assessed for similarity based on dice score, mean distance-to-agreement, and total segmented volume. Further, the contours were applied to voxel dose maps computed from post-treatment quantitative SPECT imaging to estimate renal radiation dose from therapy.

Results: Neural network segmentation was able to identify right and left kidneys in all patients with a high degree of accuracy. The system was integrated into the hospital image database, returning contours for a selected study in approximately 90 s. Mean dice score was 0.91 and 0.86 for right and left kidneys, respectively. Poor performance was observed in three patients with cystic kidneys of which only few were included in the training data. No significant difference in mean radiation absorbed dose was observed between the manual and automated algorithms.

Conclusion: Automated contouring using CNNs shows promise in providing quantitative assessment of functional SPECT and possibly PET images; in this case demonstrating comparable accuracy for radiation dose interpretation in unsealed source therapy relative to a human observer.

Keywords: automated segmentation, radionuclide therapy, kidney, nuclear medicine dosimetry, deep learning

OPEN ACCESS

Edited by:
Jun Deng,
Yale University,
United States

Reviewed by:
Seong Ki Mun,
Virginia Tech,
United States
Sunyoung Jang,
Princeton Radiation Oncology,
United States

****Correspondence:***
Price Jackson
price.jackson@petermac.org

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *29 January 2018*
Accepted: *25 May 2018*
Published: *14 June 2018*

Citation:
Jackson P, Hardcastle N, Dawe N, Kron T, Hofman MS and Hicks RJ (2018) Deep Learning Renal Segmentation for Fully Automated Radiation Dose Estimation in Unsealed Source Therapy. Front. Oncol. 8:215. doi: 10.3389/fonc.2018.00215

INTRODUCTION

In comparison to other radiation oncology modalities, personalized dosimetry assessment in unsealed source therapies is relatively uncommon. The process involves the measurement of regional uptake and pharmacokinetics followed by some calculation of radiation transport (1). In the first stage, the concentration of radiopharmaceutical is assessed on imaging and—by collecting a time series or applying known uptake and clearance parameters—an estimate of the number of disintegrations in each tissue is obtained. Finally, decays are converted into radiation absorbed dose through published self- and cross-dose factors or Monte Carlo simulation. Time-activity curve fitting by either least squares or analytical methods is a mechanical process. Similarly, integration of pharmacokinetic data and multiplication of organ or voxel dose factors are trivial mathematical operations. Unfortunately, employing these techniques often requires manual input with a degree of time and expertise that precludes their widespread use. In a previous work, we have demonstrated the feasibility of performing image-based dosimetry to create three-dimensional voxel dose maps (2). This is an automated process that may be applied to any radionuclide treatment where sequential follow-up imaging is available.

The use of neural networks for organ recognition has rapidly surpassed the capabilities of existing automated contouring techniques that rely on either rule-based methods (3) or atlas segmentation (4). Within just a few years the road map for performing pixel-by-pixel segmentation from a practical amount of ground truth data has demonstrated applications across most medical imaging modalities (5–7). These convolutional neural networks (CNNs) are demonstrating utility for image segmentation in CT, MRI, and ultrasound (8, 9). They may be designed to operate based on two-, three-, or even four-dimensional (either time series or multiparametric) images (10, 11). They have shown applications in rapid contouring to offer more efficient radiation therapy treatment planning (12) as well as in the field of computer-aided detection of specific pathologies (13). Moreover, these computational techniques—both inference and model training—are feasible on standard personal computers.

Segmentation of kidney on CT imaging presents challenges because the appearance, particularly at the inferior- and superior-most slices, may closely resemble other abdominal structures in terms of shape and physical density. As such, it is logical to employ a CNN that utilizes 3D kernels across the input volume as a whole (14). The predicted shape on one slice is then informed by features present on subsequent image slices. In this work we employ an automated CNN-based software tool to perform quantitative analysis of SPECT images based on the anatomical outline in a fused CT volume. More specifically, we demonstrate the feasibility of fully automated radiation dose estimation in unsealed source therapy as applied to patients with metastatic prostate cancer treated with radioactive prostate-specific membrane antigen (PSMA).

MATERIALS AND METHODS

Training Image Data

Training cohort was based on a population of manually contoured left and right kidneys from varied group of clinical cases. The largest of these was a set of post-treatment ^{177}Lu-octreotate therapy of neuroendocrine cancer acquired on a hybrid SPECT system with low-dose CT acquisition and 5 mm slice thickness (Siemens Symbia T6 & Intevo 16, Siemens Healthineers, Erlangen, Germany). A subset of patients scanned on dedicated diagnostic CT (Siemens Force, 0.8–5.0 mm slice thickness) and radiotherapy simulation CT systems (Philips Brilliance Big Bore, 3 mm slice thickness, Philips Medical Systems, Cleveland, OH, USA) were included to better adapt the model for detection across different populations and equipment types. A total of 89 manually contoured patients were included for training. Each patient was augmented seven times with a random degree of added noise, edge enhancement, Gaussian smoothing, change in global HU values, translation, and in-plane rotation to avoid CNN overfitting due to non-anatomical image feature (6). This provided 712 subjects available for model training. A detailed description of the image augmentation techniques used is given in the Appendix S1 in Supplementary Material.

Testing Image Data

Independent test images were taken from a cohort of 24 patients involved in a Phase II prospective trial of ^{177}Lu-PSMA-617 for treatment of metastatic prostate cancer (ANZCTR12615000912583) (15). Each patient received serial post-treatment quantitative SPECT/CT imaging (16) at timepoints of 4, 24, and 96 h. Three-dimensional radiation dose maps were processed using a previously described technique involving non-rigid image registration, voxel-wise pharmacokinetics analysis, and dose kernel convolution (2). Low-dose, fused CT images were designated as input to the CNN segmentation model. Each kidney in the testing cohort was manually contoured and reviewed by a nuclear medicine physician. Structures were compared to those automatically detected based on dice score, mean distance-to-agreement (per voxel the shortest distance from the surface of one structure to another), volume, and estimated radiation absorbed dose from ^{177}Lu therapy according to three-dimensional voxel dose map (17). Mean right and left kidney doses were evaluated for null hypothesis of difference between contour techniques by paired *t*-test.

Convolutional Neural Network

Three-dimensional convolutional neural network was modified from the structure published by Pazhitnykh et al. using 21 convolutional layers (18). CNN architecture was employed with Keras (v2.08) in Python with Tensorflow backend (v1.3) (19). A dice coefficient loss function—the ratio of the intersection of predicted and true labels over their average volume—was used to improve sensitivity to structure margins and normalize the weight of each classification region: left kidney, right kidney, and background. Each convolution layer utilizes filters with dimensions of $3 \times 3 \times 3$ followed by batch normalization (20) and rectified linear unit activation layers (21). Following convolution at each resolution a $2 \times 2 \times 2$ max pooling layer was used to downsample deeper network layers. After four convolution, normalization, activation, and max pooling stages, the network employs a similar process to upsample the native image resolution. The output of the activation layers prior to max pooling are concatenated with the output

of the upsampled activation values of the same resolution using the U-Net methodology described by Ronneberger et al. (6). The overall network framework is given in **Figure 1**.

Convolutional neural network input volume is a matrix with dimensions 128 × 128 × 64 voxels. The workflow involved several pre-processing steps. First, bony anatomy was aligned with a reference patient by rigid registration (22). Images were cropped to a smaller search volume of 334 × mm x 334 × mm x 320 mm; a volume that could consistently capture the variation in kidney location between patients, while limiting the degree of down-sampling required for input into the CNN algorithm. The native hybrid CT voxel resolution of 0.98 mm × 0.98 mm × 5.0 mm was subsequently resampled at 2.61 mm × 2.61 mm × 5.0 mm to achieve the required matrix dimensions. The complete workflow is illustrated in **Figure 2**. All training patients were pre-processed by the same methodology. Network training was allowed to run for 300 epochs using 640 teaching subjects. Another 72 augmented samples were used as a semi-independent scoring set to test training progress. Processing required 2.5 days on a cuda-enabled GPU (Nvidia GeForce GTX 1080 Ti) achieving dice accuracy of 0.98 with training data and 0.93 with a subset of augmented training patients as shown in **Figure 3**.

The software tool was integrated with the hospital PACS image database allowing selected CT studies to be transferred to a processing dicom node—a local computer—which returned the label map as a corresponding dicom-RT structure set. Structures could be viewed and modified on a standard imaging workstation and accessible hospital-wide. The process typically completed in less than 90 s; most of which was required for registration to the reference volume and post-processing to upsample the detected kidney labels at the original CT image resolution. CNNs contour detection required 10–15 s in most cases.

RESULTS

A deep learning segmentation model was trained for detection and accurate delineation of kidneys on non-contrast, low-dose CT scans. A typical result overlaid on fused CT and voxel dose map is given in **Figure 4**. In more than 80% of cases, margins were in close visual agreement for both kidneys. Visual results of manual and automated contours overlaid with a coronal maximum intensity projection of the voxel dose map for each patient are shown in **Figure 5**. Even in poorly performing cases, some region of each kidney was detected with the developed registration and CNN method; a volume that was often representative of radionuclide uptake across the organ's functional structure. When compared to manual segmentation as ground truth, automated contours achieved mean dice scores of 0.91 ± 0.05 and 0.86 ± 0.18 for right

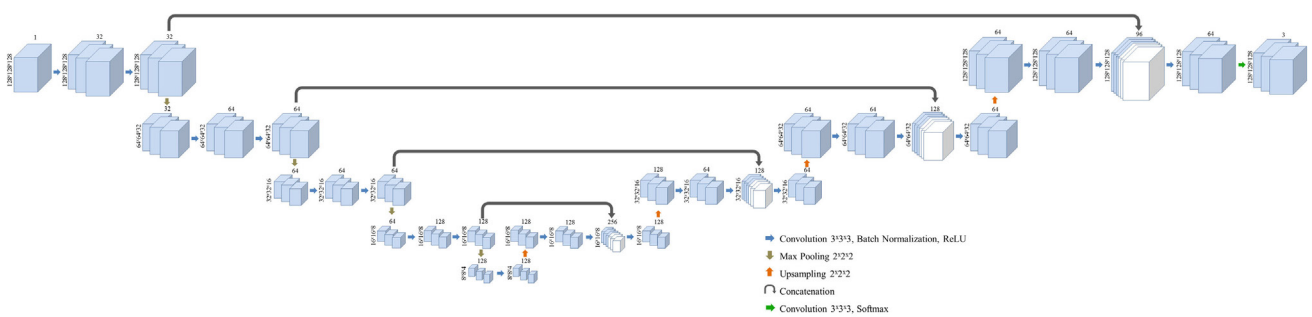

FIGURE 1 | Layer structure and dimensions for the semantic convolutional neural network used in this study. Fully three-dimensional convolutional network structure is adapted from a work by Pazhitnykh et al. (18).

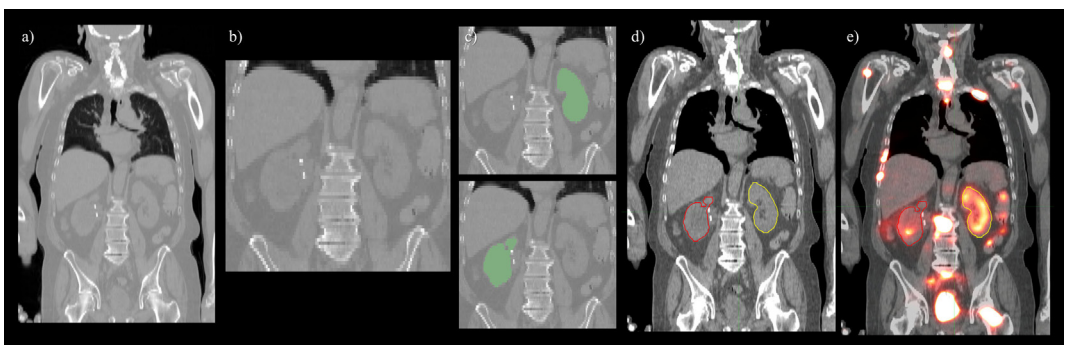

FIGURE 2 | Pre- and post-processing workflow developed for automated renal dosimetry. Input image **(A)** is automatically cropped to a smaller search area based on alignment to a reference image volume. The cropped and downsampled image **(B)**, 128 × 128 × 64 voxels, is used as input to the convolutional neural networks segmentation model. Labeled left and right kidneys **(C)** are then upsampled, smoothed, and fused with the original, uncropped image **(D)**. Label map is exported in dicom-RT structure file, where voxel dose images may be analyzed on a hospital workstation **(E)**.

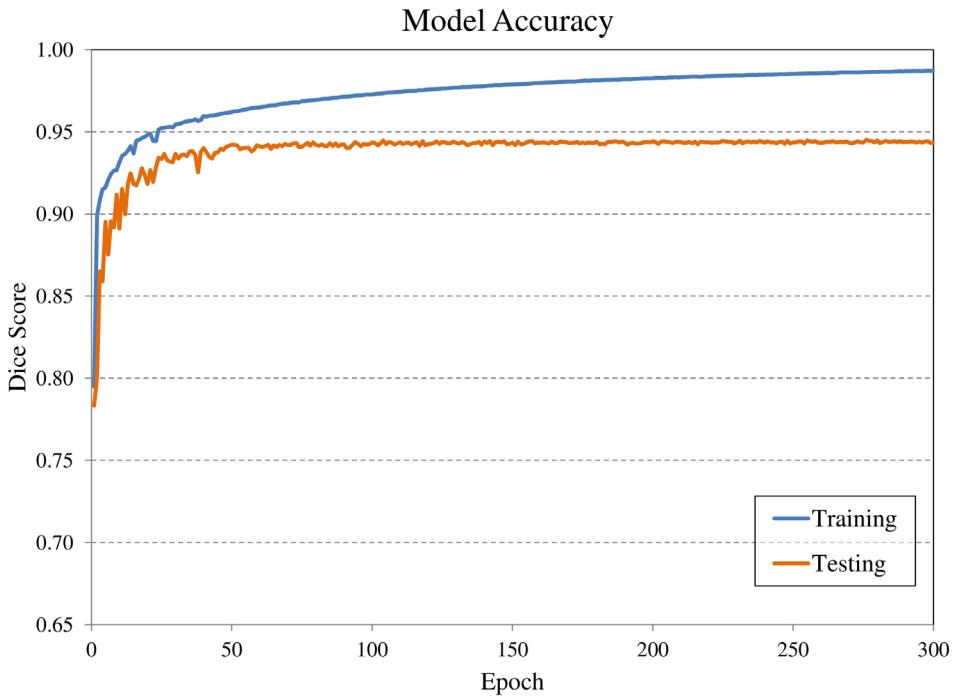

FIGURE 3 | Evolution of model accuracy with over the 300 epochs for the training cohort (blue) and an augmented testing cohort (orange). Training required approximately 2.5 days for completion.

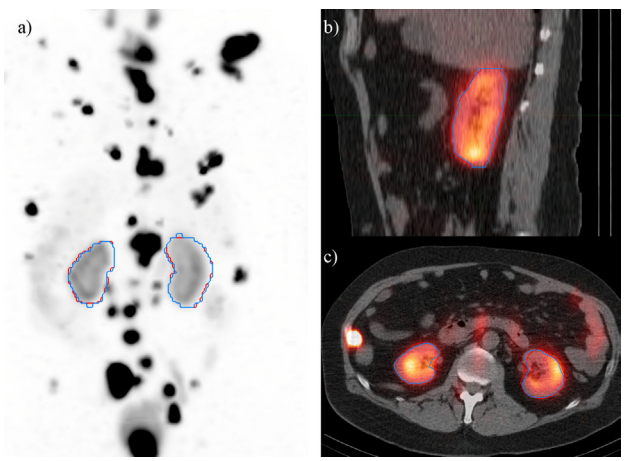

FIGURE 4 | Typical case illustrating accuracy of the deep learning segmentation algorithm (red) in comparison to manual contours (blue). Contours are shown on **(A)** maximum intensity projection of voxel dose map, as well as **(B)** sagittal and **(C)** axial fused image sets.

and left kidneys, respectively. The mean distance-to-agreement was estimated at 2.0 ± 1.0 and 4.0 ± 7.5 mm; a finer accuracy than the system resolution of typical SPECT imaging device.

Ignoring the one poorly performing left kidney with dice score of 0.11 and mean distance-to-agreement of 38.3 mm, left kidney accuracy is compared to the right side with a mean dice value of 0.89 ± 0.08 and MDA of 2.5 ± 1.7 mm. It should also be noted that the CNN-defined contours were consistently larger than those drawn manually by a factor of approximately 7%. This systematic effect likely attributed to the upsampling and smoothing of the predicted contours when returning to the native CT resolution and may be corrected by adjusting the prediction threshold to a value slightly above 0.5.

Comparing radiation dose estimates from automated and manually drawn contours, there is no apparent bias using either technique (**Figure 6**). Across the cohort there was an average difference in dose estimate of 3.0% in the right kidney and −3.6% in the left. SD of the error was ±4.5 and ±5.7%, respectively. If omitting the results for patients with cystic kidneys which would be reviewed and corrected in a clinical workflow—patients #2, 3, and 5 in **Figure 5**—the discrepancy in dose estimates between manual and automated methods is less than 2% for both kidneys. Based on *t*-test of null hypothesis, no difference between dose estimates between groups was detected ($p = 0.03$ and $p = 0.01$, right and left). Results of contour accuracy and renal radiation dose for each patient are reported in Table S1 in Supplementary Material.

Three of the patients in the ^{177}Lu-PSMA therapy cohort displayed highly cystic kidneys; to a degree that was not observed in the training patients (**Figure 7**). In these cases, the mean dice score was dramatically lower at 0.66. No systematic increase or decrease in estimated dose was shown (−2.70%) indicating that often the CNN-contoured region was representative of the mean uptake in the manually delineated kidney. In another three patients, a small, detached section (<10 cc) of bowel was included one of the contours.

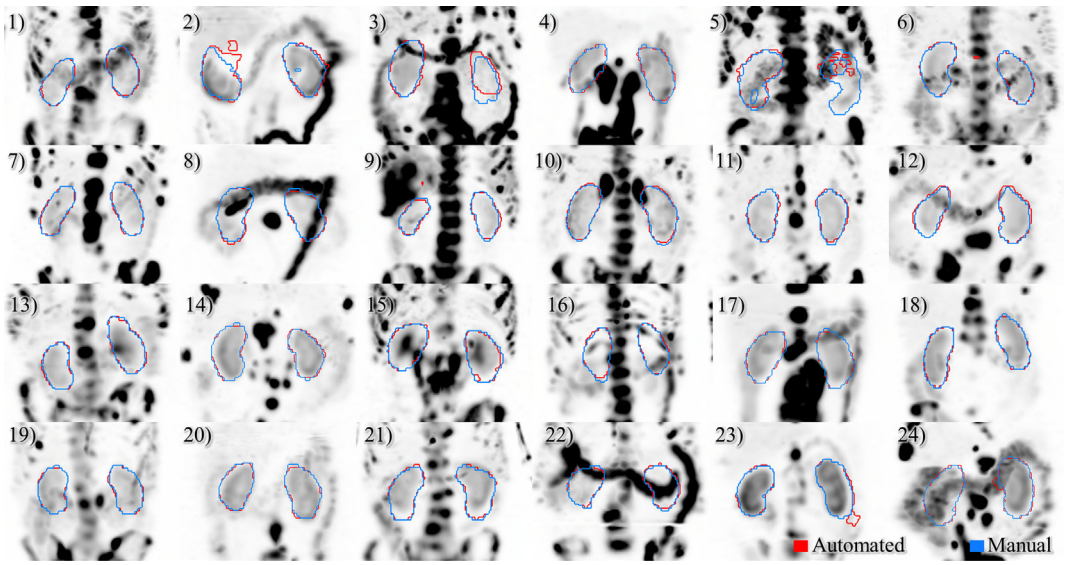

FIGURE 5 | Results of automated (red) and manual (blue) segmentation overlaid with maximum intensity projections of voxel dose volumes. Patients 2 and 5 show some disagreement in cystic regions of kidney. In patients 6, 23, and 24 a small volume of bowel is captured by the automated algorithm.

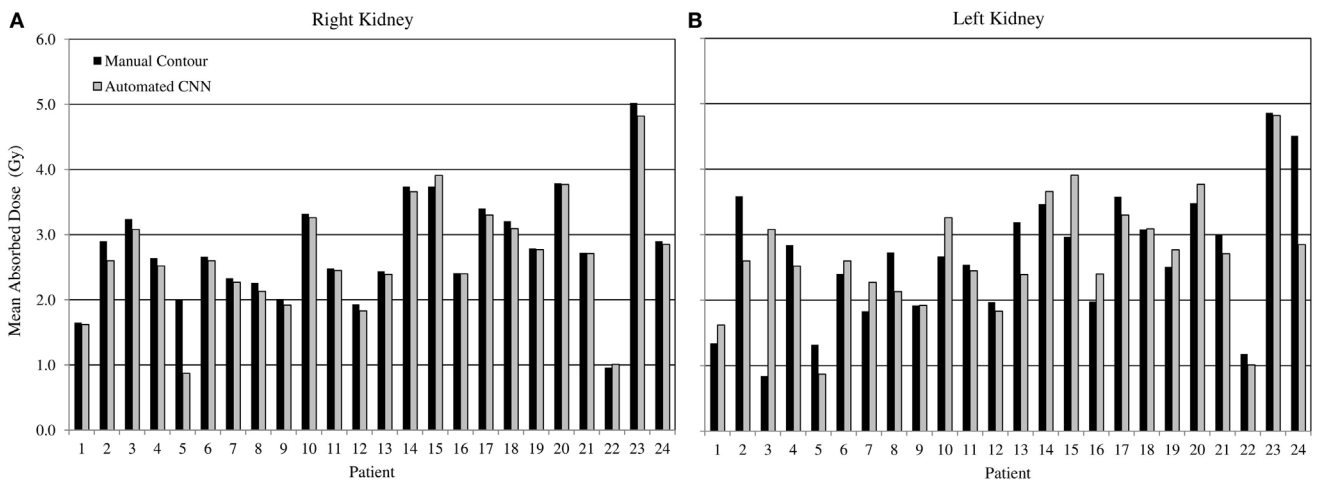

FIGURE 6 | Estimated radiation absorbed dose applying either manual (black) or automated convolutional neural network segmentation (grey) to determine mean radiation absorbed to right and left kidneys from 3D voxel dose maps. Results for right **(A)** and left **(B)** kidneys are presented independently.

In none of the cases did error manifest in an appreciable effect on estimated renal dose. If frequently noted, small non-contiguous labels could be detected and removed as a post-processing step. Only one patient with structurally normal kidneys showed poor performance with the segmentation algorithm omitting approximately one-third of the left kidney volume (dice = 0.67); an error which coincided with a region of CT streak artifact.

DISCUSSION

The advent or rapid, accurate tissue contouring through deep learning segmentation demonstrates the potential for quantitative diagnosis in molecular imaging. In this study, the results of a CNN trained to detect kidneys on CT images have been used to assess regional radiation exposure in unsealed source therapy. In principle, contouring of tumors and at-risk tissues is the last remaining step in nuclear medicine dosimetry that required manual oversight. We have combined automated kidney segmentation with a previous work that computed voxel dose maps from serial post-treatment SPECT images to demonstrate the feasibility of a fully automated system. The time required to process a case with manual methods may require several hours and may be subjected to systematic variability due to the method of curve fitting and drawn contour margins.

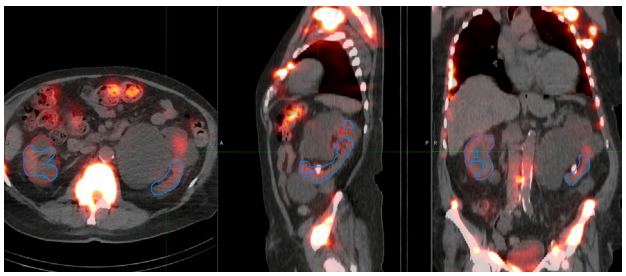

FIGURE 7 | Most challenging case encountered in testing the renal convolutional neural networks. Due to multiple large cysts in originating within the central renal structure, the segmentation tool detected only 20 cc of the manually contoured 167 cc left kidney volume (dice = 0.11).

While the automated system performs well in most cases—achieving dice scores which are comparable to inter-observer variability between manual scores in CT (23)—it is advisable to review all contours before being relied upon for quantitative assessment. In this instance, the training cohort was not necessarily representative of the patients used for testing. Those used to train the model were generally younger, from both genders, and did not include cases with cystic kidneys which were observed in 3 of the 24 testing cases. In the preparation of this framework, considerable improvement was noted over multiple iterations of the renal CNN as challenging cases were flagged, manually contoured, and incorporated into subsequent training files. It is worth noting that the addition of these irregular patients did not hinder the accuracy of the CNN when detecting otherwise normal anatomy. From the experience in developing this tool, the authors speculate that features which would accommodate detection of functional regions in polycystic kidneys would develop as the model which was retrained with additional poorly performing cases.

Previous methods such as the one described by Hasegawa do appear sound for segmentation of two-dimensional images (24). The majority of recent publications involving semantic segmentation employ variations on the U-Net structure described by Ronneberger et al. (6). These have been adapted to 3D image volumes and have proven sufficiently accurate to avoid the need for shape-based post-processing. The depth of these networks may be considered overkill when comparing the complexity of the segmentation task relative to the number of parameters that define the model weights. However, the computational requirements to train and apply such a model are feasible on a standard PC and they (or slight variations) have been shown to be extremely adaptable to a multitude of image segmentation tasks (13, 14, 25, 26). In the present work, we have, therefore, chosen to adopt the CNN approach given that a more complex algorithm may also prove more adaptable with issues concerning some of the more complex structural abnormalities, such as renal cysts.

By employing a dice score loss function based on the accuracy of trained kidney margins rather than the total number of correctly categorized voxels, a dramatic improvement in the detection of kidney margins was observed. In the former version, as employed by Pazhitnykh et al. to contour lungs (18), the model was heavily weighted to correctly designate background (non-label) voxels which typically comprised more than 90% of the search volume. In this initial iteration, the CNN could be trained to routinely identify some or the majority of kidney tissue, but was not sensitive to small boundary errors because these only manifest in subtle changes to the overall accuracy calculation. The combination of dice score and training data augmentation greatly improved the algorithm utility; correctly identifying organ margins in approximately 80% of cases. The model reported in this work was further improved by the addition of challenging cases that were flagged as poorly delineated by the existing CNN. This method could be applied to other challenging soft tissue regions and hope to implement a more comprehensive set of organs in future nuclear medicine dosimetry tools. For smaller organs or tumors, it may be advisable to utilize a tighter search volume or sliding window technique to perform classification at or near the native CT image resolution. There is also the potential to feed the fused SPECT/CT or PET/CT dataset into the CNN, capitalizing on complimentary features in both image domains to improve specificity.

CONCLUSION

Medical image segmentation by CNNs shows merit in the analysis of post-treatment scans in order to practically estimate radiation dose from unsealed source therapies. Deep learning methods have been applied to consistently detect right and left kidneys with no significant difference between radiation dose determined from CNN contours compared with manual methods. The tool has been combined with a previously developed voxel dose processing technique demonstrating the potential for fully automated radiation dose estimation for nuclear medicine therapies in the near future.

AVAILABILITY OF DATA AND MATERIALS

The patient datasets used in this study are not available to the public. The neural network model as well as pre- and post-processing computer software may be distributed on request to the corresponding author.

ETHICS STATEMENT

^{177}Lu-PMSA-617 trial was approved by the institutional ethics board and registered with the Australian New Zealand Clinical Trials Registry (ANZCTR12615000912583). The study protocol was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice and all patients gave written informed consent prior to entry on the study.

AUTHOR CONTRIBUTIONS

PJ developed the image processing techniques described in this research article. NH assisted with the similarity analysis of contour shape and radiation absorbed dose comparison. ND provided assistance with development of neural network software in Python. TK helped with study design and guided practical implementation as a hospital tool. MH was clinical lead on 177Lu-PSMA therapy trial and with the assistance of RH provided access to validation

images used in this study. All authors contributed to the review and authorization of the paper.

FUNDING

^{177}Lu (no carrier added) was supplied by the Australian National Nuclear Science and Technology Organisation (ANSTO) and PSMA-617 by Advanced Biochemical Compounds (ABX, Radeberg, Germany). MH is supported by a Clinical Fellowship Award from the Peter MacCallum Foundation and a Movember Clinical Trials Award awarded through the Prostate Cancer Foundation of Australia's Research Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fonc.2018.00215/full#supplementary-material.

REFERENCES

1. Stabin MG, Sparks RB, Crowe E. OLINDA/EXM: the second-generation personal computer software for internal dose assessment in nuclear medicine. *J Nucl Med* (2005) 46:1023–7.
2. Jackson PA, Beauregard JM, Hofman MS, Kron T, Hogg A, Hicks RJ. An automated voxelized dosimetry tool for radionuclide therapy based on serial quantitative SPECT/CT imaging. *Med Phys* (2013) 40:112503. doi:10.1118/1.4824318
3. Massoptier L, Casciaro S. Fully automatic liver segmentation through graph-cut technique. *Engineering in Medicine and Biology Society, 2007 EMBS 2007 29th Annual International Conference of the IEEE*. Lyon, France: IEEE (2007). p. 5243–6.
4. Gorthi S, Duay V, Houhou N, Cuadra MB, Schick U, Becker M, et al. Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. *IEEE J Sel Top Signal Process* (2009) 3:135–47. doi:10.1109/JSTSP.2008.2011104
5. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston (2015). p. 3431–40.
6. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich: Springer (2015). p. 234–41.
7. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans Med Imaging* (2016) 35:1153–9. doi:10.1109/TMI.2016.2553401
8. Men K, Dai J, Li Y. Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. *Med Phys* (2017) 44:6377–89. doi:10.1002/mp.12602
9. Cheng J-Z, Ni D, Chou Y-H, Qin J, Tiu C-M, Chang Y-C, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci Rep* (2016) 6:24454. doi:10.1038/srep24454
10. Shin H-C, Orton MR, Collins DJ, Doran SJ, Leach MO. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Trans Pattern Anal Mach Intell* (2013) 35:1930–43. doi:10.1109/TPAMI.2012.277
11. Tsehay YK, Lay NS, Roth HR, Wang X, Kwaka JT, Turkbey BI, et al. Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images. *SPIE Medical Imaging: International Society for Optics and Photonics*. Orlando, FL (2017). p. 1013405–11.
12. Sun W. *Deep Learning Method vs. Hand-Crafted Features for Lung Cancer Diagnosis and Breast Cancer Risk Analysis*. El Paso, TX: The University of Texas at El Paso (2017).
13. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* (2016) 35:1285–98. doi:10.1109/TMI.2016.2528162
14. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: learning dense volumetric segmentation from sparse annotation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Athens: Springer (2016). p. 424–32.
15. Hofman MS, Sandhu S, Eu P, Price J, Akhurst T, Iravani A, et al. 785OLutetium-177 PSMA (LuPSMA) theranostics phase II trial: efficacy, safety and QoL in patients with castrate-resistant prostate cancer treated with LuPSMA. *Ann Oncol* (2017) 28(Suppl 5). doi:10.1093/annonc/mdx370.002
16. Beauregard J-M, Hofman MS, Pereira JM, Eu P, Hicks RJ. Quantitative 177Lu SPECT (QSPECT) imaging using a commercially available SPECT/CT system. *Cancer Imaging* (2011) 11:56. doi:10.1102/1470-7330.2011.0012
17. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol* (2016) 121:169–79. doi:10.1016/j.radonc.2016.09.009
18. Pazhitnykh I, Petsiuk V. *Lung Segmentation (3D)*. (2017). Available from: https://github.com/imlab-uiip/lung-segmentation-3d (Accessed: 1 August, 2017).
19. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:160304467*. (2016).
20. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:150203167*. (2015).
21. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Haifa (2010). p. 807–14.
22. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of SimpleITK. *Front Neuroinformatics* (2013) 7:45. doi:10.3389/fninf.2013.00045
23. Shim H, Chang S, Tao C, Wang JH, Kaya D, Bae KT. Semiautomated segmentation of kidney from high-resolution multidetector computed tomography images using a graph-cuts technique. *J Comput Assist Tomogr* (2009) 33:893–901. doi:10.1097/RCT.0b013e3181a5cc16
24. Hasegawa A, Lo S-CB, Lin J-S, Freedman MT, Mun SK. A shift-invariant neural network for the lung field segmentation in chest radiography. *J VLSI Signal Process Syst Signal Image Video Technol* (1998) 18:241–50. doi:10.1023/A:1007937214367
25. Roth HR, Farag A, Lu L, Turkbey EB, Summers RM. Deep convolutional networks for pancreas segmentation in CT imaging. *Medical Imaging 2015: Image Processing: International Society for Optics and Photonics*. Orlando (2015). 94131G p.
26. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. *International Workshop on Machine Learning in Medical Imaging*. Quebec City: Springer (2017). p. 379–87.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

REVIEW
published: 21 June 2018
doi: 10.3389/fonc.2018.00228

Machine Learning and Radiogenomics: Lessons Learned and Future Directions

John Kang[1], Tiziana Rancati[2], Sangkyu Lee[3], Jung Hun Oh[3], Sarah L. Kerns[1], Jacob G. Scott[4,5], Russell Schwartz[6,7], Seyoung Kim[6] and Barry S. Rosenstein[8,9]*

[1] Department of Radiation Oncology, University of Rochester Medical Center, Rochester, NY, United States, [2] Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [3] Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [4] Department of Translational Hematology and Oncology Research, Cleveland Clinic, Cleveland, OH, United States, [5] Department of Radiation Oncology, Cleveland Clinic, Cleveland, OH, United States, [6] Computational Biology Department, Carnegie Mellon School of Computer Science, Pittsburgh, PA, United States, [7] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, United States, [8] Department of Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [9] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Due to the rapid increase in the availability of patient data, there is significant interest in precision medicine that could facilitate the development of a personalized treatment plan for each patient on an individual basis. Radiation oncology is particularly suited for predictive machine learning (ML) models due to the enormous amount of diagnostic data used as input and therapeutic data generated as output. An emerging field in precision radiation oncology that can take advantage of ML approaches is radiogenomics, which is the study of the impact of genomic variations on the sensitivity of normal and tumor tissue to radiation. Currently, patients undergoing radiotherapy are treated using uniform dose constraints specific to the tumor and surrounding normal tissues. This is suboptimal in many ways. First, the dose that can be delivered to the target volume may be insufficient for control but is constrained by the surrounding normal tissue, as dose escalation can lead to significant morbidity and rare. Second, two patients with nearly identical dose distributions can have substantially different acute and late toxicities, resulting in lengthy treatment breaks and suboptimal control, or chronic morbidities leading to poor quality of life. Despite significant advances in radiogenomics, the magnitude of the genetic contribution to radiation response far exceeds our current understanding of individual risk variants. In the field of genomics, ML methods are being used to extract harder-to-detect knowledge, but these methods have yet to fully penetrate radiogenomics. Hence, the goal of this publication is to provide an overview of ML as it applies to radiogenomics. We begin with a brief history of radiogenomics and its relationship to precision medicine. We then introduce ML and compare it to statistical hypothesis testing to reflect on shared lessons and to avoid common pitfalls. Current ML approaches to genome-wide association studies are examined. The application of ML specifically to radiogenomics is next presented. We end with important lessons for the proper integration of ML into radiogenomics.

Keywords: statistical genetics and genomics, radiation oncology, computational genomics, precision oncology, machine learning in radiation oncology, big data, predictive modeling

OPEN ACCESS

Edited by:
Jun Deng,
Yale University, United States

Reviewed by:
Jonathan W. Lischalk,
Georgetown University, United States
Sunyoung Jang,
Princeton Radiation Oncology, United States

****Correspondence:***
John Kang
johnkan1@alumni.cmu.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *16 February 2018*
Accepted: *04 June 2018*
Published: *21 June 2018*

Citation:
Kang J, Rancati T, Lee S, Oh JH, Kerns SL, Scott JG, Schwartz R, Kim S and Rosenstein BS (2018) Machine Learning and Radiogenomics: Lessons Learned and Future Directions. Front. Oncol. 8:228. doi: 10.3389/fonc.2018.00228

1. INTRODUCTION TO RADIOGENOMICS

1.1. Normal Tissue Toxicity Directly Limits Tumor Control

Over 50 years before the discovery of the DNA double helix, radiation therapy and normal tissue radiobiology became irrevocably linked after Antoine Henri Becquerel left a container of radium in his vest pocket, causing a burn-like reaction of erythema followed by ulceration and necrosis (1, 2). Ever since, the goal of therapeutic radiation has been to deliver a maximal effective dose while minimizing toxicity to normal tissues. The importance of this goal has increased as cancers that were previously fatal became curable and patients have had to live with long-lasting late effects and secondary malignancies (3, 4).

For several tumors, an argument can be made that survival is so poor that one should not be as concerned for late effects. However, acute toxicity may also constrain dose escalation, which directly limits tumor control, since a therapeutically efficacious dose may not be achievable due to toxicity. This is because dose tolerances are typically set for 5–10% toxicity in clinical trials, so the patients with the most radiosensitive normal tissue ultimately determine the limit for the maximum dosage for all patients (5, 6). As Becquerel noted, tumor control and normal tissue toxicity have been, and remain, irrevocably linked. Advances in the last decades from the fields of radiation physics and radiation biology have focused on finding ways to separate these two effects with varying success, as discussed below.

1.2. Technology Has Improved Normal Tissue Toxicity

To improve therapeutic ratio (i.e., the cost–benefit of tumor control vs. normal tissue side effects) in recent decades, medical physics has made significant advances in the technology and techniques of radiation delivery to spare normal tissue (7). This includes moving from 2D treatment planning using X-ray films to 3D planning using CT-simulation, and now to inverse planning and fluence modulation to create conformal dose distributions employing intensity-modulated radiation therapy (IMRT) (8). IMRT not only utilizes more sophisticated hardware but also advanced treatment planning software and optimization algorithms. Multiple prospective and retrospective studies have demonstrated the superiority of IMRT in reducing toxicity for most solid cancer types, including those of the head and neck (9), lung (10), prostate (11), anus (8), and soft tissue sarcoma (12). Utilizing protons for cancer treatment provides another way to increase dose conformality and decrease normal tissue dose through the Bragg peak. Complementary technologies include improvements in image guidance (13), motion management (14), and patient positioning (15). Radiosurgery for central nervous system tumors is an attractive alternative to lengthier and more toxic treatments. Brachytherapy also offers dosimetric advantages to decrease toxicity and improve tumor control. Due to the successes of the technological advancements, there has been relatively fast adoption of emerging physics technologies in the clinic as standard of care in many places.

1.3. Radiobiology and Normal Tissue Toxicity

While radiation physics was using increasingly complex methods and data to perform more individualized treatments, advancements in radiation biology were also developing, but have yet to achieve the same level of clinical impact. Early efforts in the 1980s and 1990s to employ radiation biology approaches in the clinic focused on altered fractionation schedules to improve control of head and neck tumors and small cell lung cancer while sparing normal tissue toxicity. These trials demonstrated benefits to both hyperfractionation (16, 17) and accelerated fractionation (18, 19), but these protocols have not translated into changes in the standard of care at many centers or into similar studies in most cancers (20). Therapies for modulating tissue oxygenation and the use of hypoxic cell radiosensitizers and bioreductive drugs have been moderately successful in animal studies and randomized clinical trials (21) but also have not yet reached wide penetration in the United States despite level I evidence, often due to side effects. More recently, hypofractionation (i.e., larger doses of radiation per fraction) has become widely adopted; however, there is significant controversy as to how this can best be modeled (22–26). Whereas advances in radiation physics brought about measurable improvements in both tumor control and normal tissue protection as demonstrated through multiple clinical trials—largely due to IMRT—this could not be said for advances in radiobiology. It became clear that a different approach other than modeling of fractionation would be necessary to keep pace with the increasing torrent of clinical data. Such an opportunity would arise at the turn of the twenty-first century with substantial advances in molecular biology and the first draft of the human genome (27, 28) as discussed below.

1.4. Genomic Basis for Radiotherapy Response

Through studies of patients following radiotherapy (29, 30), it has become apparent that patient-related characteristics, including genomic factors, could influence susceptibility for the development of radiation-related toxicities (31). To identify the genomic factors that may be associated with normal tissue toxicities, a series of candidate gene studies was performed that resulted in more than 100 publications from 1997 to 2015 (32). However, with a few exceptions, the findings were largely inconclusive, and independent validations were rare (33). The risk of spurious single-nucleotide polymorphism (SNP) associations has been a concern for candidate gene association studies even before the advent of genome-wide association studies (GWAS) (34).

With improved understanding of the genetic architecture of complex traits, we now know that a few variants in limited pathways—such as DNA damage response—cannot alone explain most of variation in radiotherapy response. While this work was in progress, results of the Human Genome Project and related efforts demonstrated the magnitude of genetic variation between individuals. Over 90% of this variation comes from common SNPs (frequency >1%) and rare variants. There are about 10 million common SNPs in the human genome and any locus can be affected. These variants can be in coding regions (exons), introns,

or intergenic regulatory regions. Early efforts to understand how SNPs were linked to phenotypic traits were marred by poor statistical understanding of correction for multiple hypothesis testing, which led to multiple small and underpowered studies (35).

To improve power to detect new SNP biomarkers for radiation toxicity, the International Radiogenomics Consortium (RGC) was formed in 2009 to pool individual cohorts and research groups. One of the main goals is to determine germline predisposition to radiation toxicity and there have been several studies from RGC investigators that have identified novel risk SNPs.

REQUITE is a project led by RGC members to prospectively collect clinical and biological data, and genetic information for 5,300 lung, prostate, and breast cancer patients (36). The RGC also collaborates with the GAME-ON oncoarray initiative (32).

1.4.1. Fundamental Hypothesis of Radiogenomics

Andreassen et al. reported three basic hypotheses of radiogenomics (32):

(a) Normal tissue radiosensitivity is as a complex trait dependent on the combined influence of sequence alteration of several genes.
(b) SNPs may make up a proportion of the genetics underlying differences in clinical normal tissue radiosensitivity.
(c) Some genetic alterations are expressed selectively through certain types of normal tissue reactions, whereas others exhibit a "global" impact on radiosensitivity.

Regarding these hypotheses, it is prudent to add that we are now aware that there are also epigenetic components of normal tissue radiosensitivity that are—by definition—not captured by genetic sequences but are heritable nonetheless.

1.4.2. The Importance of Fishing

Genome-wide association studies could certainly be categorized as a "fishing expedition," which has pejorative connotations given the history of improper correction for multiple hypothesis testing (see Multiple Hypothesis Correction). However, fishing expeditions in genomics are a necessity to generate new hypotheses. Recent GWAS performed by members of the RGC have been able to identify novel associations of SNPs in genes that were previously not linked with radiation toxicity (37). For example, *TANC1* is a gene that encodes a repair protein for muscle damage and is one such example of a novel radiosensitivity association discovered in 2014 (38). A meta-analysis of four GWAS also identified two SNPs, rs17599026 in *KDM3B* and rs27720298 in *DNAH5*, which are associated with increased urinary frequency and decreased urinary stream, respectively (39).

1.5. Precision Medicine and Single Drug Targets

Compared to biomarker panels for normal tissue toxicity to radiation therapy, the realm of biomarker panels for prediction of tumor response is a much wider field, as it also encompasses the domains of medical and surgical oncology. Early successes in predictive biomarkers focused on single mutations, such as the *BCR-ABL* translocation observed in chronic lymphocytic leukemia or oncogene amplification, such as *Her2-neu* or *EGFR*. In the last half decade, therapies targeting tyrosine kinase mutations in lung cancer or high expressing immune markers in many tissue types have become standard of care. In March 2017, the US Food and Drug Administration (FDA) granted a tissue-agnostic "blanket approval" for the PD-1 inhibitor pembrolizumab for any metastatic or unresectable solid tumor with specific mismatch repair mutations (40); this was the first time FDA approval had been granted for a specific mutation regardless of tumor type.

Given the various targeted agents, there are many who herald this as the age of "precision medicine." In late 2016, the American Society for Clinical Oncology (ASCO) launched Journal of Clinical Oncology (JCO) subjournals "JCO Clinical Cancer Informatics" and "JCO Precision Oncology." In accordance with the single target–single drug approach, contemporary precision medicine drug trials are based on amassing targetable single mutations (NCI-MATCH) or pathway mutations (NCI-MPACT) (41). While the initial tumor response can be quite impressive, durable response is an issue as single-target drugs are prone to develop resistance (42, 43).

1.6. Precision Medicine and Multigene Panels

Since the discovery of the Philadelphia chromosome and imantinib, most drugs remain focused on single biomarkers, such as a single mutation or a gene expression alteration with a large penetrance. However, we are rapidly depleting the pool of undiscovered, highly penetrant genes. Soon, targeting the low hanging fruit through a one gene–one phenotype approach will no longer be sufficient for effective "precision medicine." This is where multiple biomarker panels are making an impact. While these do not necessarily provide "multiple targets" for drugs to act on, they do provide a prognostic picture of the effects of tumor mutational burden. The earliest and most well known of these laboratory-developed biomarker panels are the 21-gene recurrence score Oncotype DX (Genomic Health, Inc., Redwood City, CA, USA) (44) and 70-gene MammaPrint (Agendia BV, The Netherlands) (45). These panels are used to make critical clinical decisions regarding whether select breast cancer patients are predicted to benefit from chemotherapy.

Current efforts are aimed at understanding the genomic signature of metastatic cancer. Memorial Sloan Kettering has used their MSK-IMPACT gene expression panel to sequence tumors from over 10,000 patients with metastatic disease to be able to prognosticate whether a future patient will develop metastases (46). While the development of these laboratory tests requires significant investment, they may ultimately save substantial sums by decreasing unnecessary therapies and toxicities while improving quality of life for cancer patients.

Recent discussions about the state of precision medicine and genomically guided radiation therapy include a review by Baumann et al. (7) and a joint report by the American Society for Radiation Oncology (ASTRO), American Association of Physicists in Medicine (AAPM), and National Cancer Institute (NCI) summarizing a 2016 precision medicine symposium (6) (see Promoting Research).

A complicating factor in tumor genomics is a result of tumor heterogeneity, which results in different subtypes within the same tumor, as shown in glioblastoma (47), colorectal cancer (48), and pancreatic cancer (49). Given the limited ability of single-target drugs, therapies may select certain subclones of higher fitness to predominate and create mechanisms of resistance. Selection occurs not only from therapy but also from local and microenvironment constraints (50), leading to an increasingly robust evolutionary model of tumor heterogeneity obeying Darwinian selection. Distant metastases display this evolutionary behavior as well as they seed further distant metastases (51). To better target a tumor's genomic landscape, we may need to sample multiple spatially separated sites and incorporate evolutionary analysis (52).

1.7. Tumor Control and Radiogenomics

Although a substantial emphasis of radiogenomics has been to identify biomarkers predictive of normal tissue toxicities, there are efforts being made to develop tests for tumor response to radiation (53). In the largest preclinical study, Yard et al. showed that there is a rich diversity of resultant mutations after exposing 533 cell lines across 26 tumor types to radiation (54). Within these tumor cell lines, radiation *sensitivity* was enriched in gene sets associated with DNA damage response, cell cycle, chromatin organization, and RNA metabolism. By contrast, radiation *resistance* was associated with cellular signaling, lipid metabolism and transport, stem-cell fate, cellular stress, and inflammation.

PORTOS is a 24-gene biomarker predictive assay that can determine which post-prostatectomy patients would benefit from post-operative radiation therapy to decrease their 10-year distant metastasis-free survival (55). PORTOS is the first of future clinical radiogenomics assays to help determine which patients will benefit from radiation.

The radiosensitivity index (RSI) was developed at Moffitt Cancer Center to predict radiation sensitivity in multiple tumor types (56, 57). Its signature is based on linear regression on the expression of 10 specific genes (*AR*, *cJun*, *STAT1*, *PKC*, *RelA*, *cABL*, *SUOMO1*, *CDK1*, *HDAC1*, and *IRF1*) that were chosen from a pool of over >7,000 genes using a pruning method derived from systems biology principles. These genes are implicated in pathways involved in DNA damage response, histone deacetylation, cell cycle, apoptosis, and proliferation. More recently, the RSI has been combined with the linear quadratic model of cell kill to create a unified model of both radiobiologic and genomic variables to predict for radiation response and provide a quantitative link from genomics to clinical dosing (58).

2. INTRODUCTION TO MACHINE LEARNING (ML)

Machine learning is a field evolved from computer science, artificial intelligence, and statistical inference that seeks to uncover patterns in data to make future predictions. Unlike handcrafted heuristic models often seen in clinical medicine, ML methods have a foundation in statistical theory and are generalizable to a type of problem as opposed to specific problems (59). There are many ML methods, and each has unique advantages and disadvantages that merit consideration by the user prior to attempting to model their results (60, 61). Similarly, there are several ML-friendly programming languages and specialized libraries to choose from, including Python's Scikit-learn package (62), MATLAB's Statistics and Machine Learning Toolbox (63), and R (64).

2.1. Statistical Inference vs. ML

Machine learning has considerable overlap with classical statistics and many key principles and methods were developed by statisticians. There continues to be considerable crossover between computer science and statistics. Breiman wrote about the differences between the two fields, calling ML the field of black box "algorithmic models" and statistics the field of inferential "data models" (65).

In ML, models are commonly validated by various measures of raw predictive performance, whereas in statistics, models are evaluated by goodness of fit to a presumptive model. These models can be used for either explaining or predicting phenomena (66). One key difference that readers of clinical papers will immediately notice is that formal hypothesis testing is a rarity in ML. This stems from the fact that ML is concerned with using prior information to improve models, rather than inferring a "belief" between two hypotheses. Classical hypothesis testing—used in most clinical studies—relies on the frequentist approach to probability. In this interpretation, one selects a level of belief α and—assuming a certain probability distribution—then determines whether the obtained result is extreme enough such that if the experiment was repeated many times, one would see this result at a rate of $\leq\alpha$. This rate is called the *p*-value, and the significance level α is typically set at 0.05. ML papers rarely discuss significance levels, instead seeking to identify maximum likelihood models or sample over spaces of possible models, as in Bayesian statistics. To determine significance levels requires some assumptions regarding the distribution implied by a null hypothesis for the data, which is more difficult for complex problems such as speech recognition, image recognition, and recommender systems.

2.2. An Update of Breiman's Lessons From ML

In 2001, Breiman noted three important lessons from ML over the prior 5 years: the Rashomon effect, the Occam dilemma, and the curse of dimensionality. Here, we will re-visit these to discuss relevance to contemporary issues of ML usage in medicine.

2.2.1. Rashomon Effect

The Rashomon effect describes a multiplicity of models where there are many "crowded" models that have very similar performance (i.e., accuracies within 0.01) but which may have very different compositions (i.e., different input variables). Within oncology, this effect is well demonstrated in breast cancer where Fan et al. showed that four of five different gene expression models (including MammaPrint and Oncotype DX Recurrence Score) showed significant agreement in patient prognosis despite having very different inputs (67). This model crowding is magnified by

variable pruning (i.e., feature selection) as the remaining variables must then *implicitly* carry the effect of the removed variables. The Rashomon effect is popularly seen in nutritional epidemiology where observational studies routinely seem to show conflicting data about the risk or benefits of certain supplements (68). This phenomenon was studied in Vitamin E, where depending on which combinations of 13 covariates were selected, one could find a range in increase or decrease of Vitamin E-associated mortality—a so-called "vibration of effects" (69).

The Rashomon effect can manifest as model instability when multiple Monte Carlo repetitions of cross-validated model selection are performed (see CV Methodology) that result in different models selected in each repetition. This occurs due to minor perturbations in the data resulting from different splits and is particularly magnified for smaller datasets. Ensemble models (70) and regularization methods (see Embedding Feature Selection With the Prediction Model) (71) seem to work well for addressing this problem.

2.2.2. Occam Dilemma

William of Occam (c. 1285–1349) described the Principle of Parsimony as: "one should not increase, beyond what is necessary, the number of entities required to explain anything." Breiman describes the Occam dilemma as the choice between simplicity—and interpretability—and accuracy. He noted that simple classifiers—such as decision trees and logistic regression (LR)—were interpretable but were easily outclassed in classification performance by more complex and less-interpretable classifiers like random forests (RFs). However, increasing model complexity also tends to overfit. This dilemma has been partially mitigated by a better understanding of cross validation (CV) (see Cross Validation) as well as better strategies for automated control for model complexity.

In contemporary usage, where the boundary between interpretable statistical models and "black box" ML models has become blurred, interpretability and accuracy discussions have resurfaced in the form of generative and discriminative models. Generative approaches resemble statistical models where the full joint distribution of features is modeled (see Bayesian Networks). Discriminative approaches focus on optimizing classification accuracy using conditional distributions to separating classes (see Support Vector Machines). Both of these approaches have been described in ML applications to genomics (72). Generative models are more interpretable and handle missing data better, whereas discriminative classifiers perform better asymptotically with larger datasets (73). Thus, we can update Breiman's interpretation with a contemporary interpretation of modeling genetic information (**Figure 1**).

Breiman had postulated that physicians would reject less-interpretable models, but this has not been the case. As discussed in Section "Precision Medicine and Multigene Panels," oncology is moving toward validating and using high-dimensional multigene models in the clinic to guide treatment decisions.

As a future where a multigene panel for all cancers is still a long way off, creating intuitive models is still relevant. Patients can rarely be placed into neat boxes, and physicians must often incorporate clinical experience, which becomes more difficult for less-interpretable models. A method that was developed to overcome this limitation is MediBoost, which attempts to emulate the performance of RF while maintaining the intuition of classic decision trees (74). In Section "Current ML Approaches to Radiogenomics," we discuss the interpretability of three ML methods.

2.2.3. The Curse of Dimensionality

The curse of dimensionality refers to the phenomenon where potential data space increases exponentially with the number of dimensions (75). For example, a cluster of points on a line of length 3 au appears much more desolate when clustered in a cube of volume 27 au^3. Two things happen with increasing dimensions: (1) available data becomes increasingly sparse and (2) the number of possible solutions increases exponentially while each can become statistically insignificant by overfitting to noise (76). Traditional thinking has always been to try to reduce feature number; however, some ML methods benefit from higher dimensions. For example, when data are nearly linearly separable, LR and linear support vector machine (SVM) perform similarly. However, when data are *not* linearly separable, SVM can use the kernel trick that increases the dimensionality of data to allow separation in higher dimension (see Support Vector

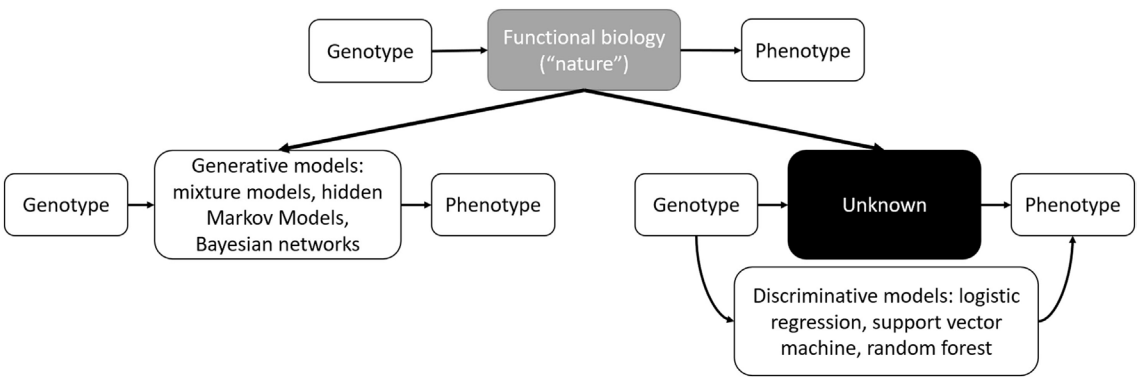

FIGURE 1 | Schematic outline of functional biology modeling *via* generative or discriminative models.

Machines). While SVM has built-in protections for this "curse" by defining kernel functions around the data points themselves and selecting only the most important support vectors, it remains vulnerable when too many support vectors are selected with high-dimensional kernels.

Within genomics, the curse of dimensionality is reflected in the difficulty of finding epistatic interactions (77). In standard search for additive genetic variance, one needs to only search n SNPs in a single dimension. However, if pairwise or higher-order interactions are considered, then the search space increases exponentially; for example, the search space for pairwise interactions is $n(n-1)/2$. Traversing the large but sparse search space while maintaining reasonable performance can be a challenge (see Combining ML and Hypothesis Testing).

2.2.4. ML Workflow

In an ideal world, there would exist a perfect protocol to follow that will guarantee a great ML model every time. Unfortunately, there is no consensus on the "optimal" way to create a model. Libbrecht and Noble described general guidelines for applying ML to genomics (72). Within radiation oncology, Lambin et al. provide a high-level overview of clinical decision support systems (78). Kang et al. discussed general ML design principles with case examples of radiotherapy toxicity prediction (60). El Naqa et al. provide a comprehensive textbook of ML in radiation oncology and medical physics (79). **Figure 2** provides a sample workflow for a general radiation oncology project that incorporates both genomics and clinical/dosimetric data. Two critical components of model selection include "Cross validation" and "Feature selection," which are further discussed below.

2.3. Cross Validation

The greater the number of parameters in a model, the better it will fit a given set of data. As datasets have become more and more complex, there has become an inherent bias toward increasing the number of parameters. Overfitting describes the phenomenon of creating an overly complex model which may fit a given data set, but will fail to generalize (i.e., fit another data set sampled from a similar population). CV is a method used in model selection aimed to prevent overfitting by estimating how well a model will generalize to unseen data.

2.3.1. CV Methodology

Conceptually, CV is used to prevent overfitting by training with data separate from validation data. As an example, in k-fold CV (KF-CV) for $k = 10$, the data are initially divided

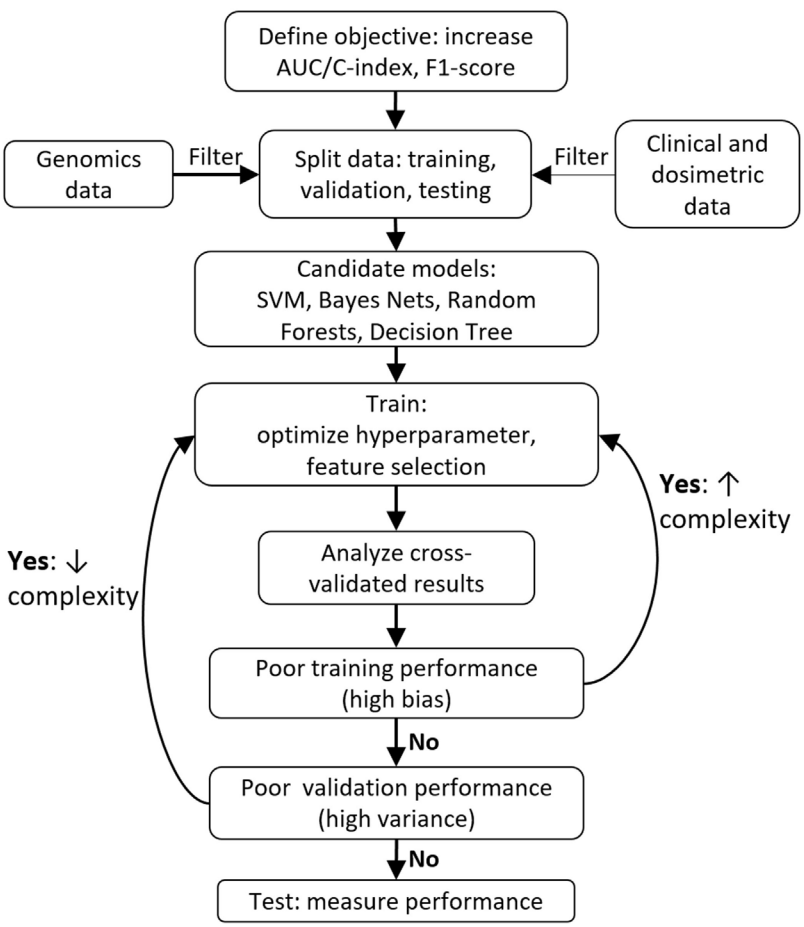

FIGURE 2 | Typical machine learning project workflow.

into 10 equal parts. Next, 9 parts are used to train a model while the 10th part is used to assess for how well the model was trained in the validation step. This training–validation procedure is run 9 more times, with each of the 10 parts taking turns as the validation set. The performance averaged over 10 runs is the cross-validated estimate of how well the model will perform on truly unseen data. The optimal number of initial splits for the data has not been established, but 10-fold CV is commonly used. An alternative to KF-CV that is often used for smaller datasets is "leave one out" cross-validation (LOO-CV), whereby a dataset of size *n* is split into *n* parts. This form of CV maximizes the relative amount of information used for training the model while minimizing the information used for testing. As a result, LOO-CV is prone to higher variance (i.e., a higher propensity to overfit) and decreased bias (i.e., a lower propensity to underfit) compared with KF-CV. Similar to balancing type I and type II error in statistical genetics, variance and bias must be carefully considered to avoid "false positive" and "false negative" results.

2.3.2. CV Relationship With Statistical Inference

Cross validation took some time to catch on in statistics literature, but has long been a fundamental part of the algorithmic ML models (65). Due to the lack of interpretability in the "black box," ML has relied on CV and related methods like bootstrapping to demonstrate robust performance without relying formally on statistical significance. Small sample sizes can be a problem for creating prediction models. In this case, learning curve analysis can be used to create empirical scaling models, whereby one varies the size of the training set to assess for learning rate (80). Learning curve analysis can be used to help determine at what point a model is overfitting (81). When learning curve analysis predicts large error rates that are unlikely to be significant, permutation testing can predict the significance of a classifier by comparing its performance with that of random classifiers trained on randomly permuted data (80).

2.4. Common Errors in CV

When performed correctly, CV is a powerful tool for selecting models that will generalize to new data. However, this seemingly simple technique is infamous for being used incorrectly. This creates an especially egregious problem as using CV gives results an appearance of rigorous methodology when the exact opposite may be occurring.

2.4.1. Violating the Independence Assumption

A common mistake is to pre-maturely "show" the test data while still training the model and thus violate the independence assumption between the training and test data. For example, a typical workflow is to set aside test data and train a model using only the training data. Once the training results are acceptable, the model is tested on the independent testing data. If the testing results are unacceptable, one might then use these results to refine the model. However, using performance on the test set to guide decisions for training, the model creates bias and violates the independence assumption between the model design and testing (82). The more repetitions of model pruning are performed, the higher the chance of the model overfitting to truly independent data. See Section "Reusable Hold-Out Set" for a solution to this problem.

Sometimes, re-using training samples in testing is intentional. This was the case in the MammaPrint assay, where the authors used a large proportion of the tumor samples from the initial discovery study in their validation study (83, 84). The authors claimed this was necessary due to an imbalance of tumor cases and controls (see Section "Unbalanced Datasets" below for solutions).

In part due to the lack of independence between the testing and training sets in biomedical research, which culminated in the pre-mature use of omics-based tests used in cancer clinical trials at Duke University (85, 86), the Institute of Medicine released a report in 2012 (84). Several cautionary steps were advised, including validating with a blinded dataset from another institution (see Replication and Regulatory Concerns).

2.4.2. Freedman's Paradox

Freedman showed that in high-dimensional data, some variables will be randomly associated with an outcome variable by chance alone and if these are selected out in model selection, they will appear to be strongly significant in an effect called Freedman's paradox (87). This can occur even with no relationship between the input variables and outcome variables because with enough input variables, by chance one will have a high correlation. Even if model selection is performed and low performing variables are removed, the same randomly associated features will remain correlated and appear to be highly significant. Freedman's paradox manifests when CV is repeated to perform both model selection and performance estimation. One solution is to use cross model validation, also known as nested CV: the outer loop is used for performance estimation and the inner loop for model selection (88–90).

2.5. Feature Selection

Often, one is interested in not only fitting an optimal model but rather in determining which of the variables—also known as features—are the most "important" through the process of feature selection. With respect to ML in genomics, Libbrecht and Noble described three ways to define "importance" in feature selection (72). The first is to identify a very small subset of features that still has excellent performance (i.e., to create a cheaper SNP array to test association with a phenotype rather than whole genome sequencing). The second is to attempt to understand underlying biology by determining which genes are the most relevant. The third is to improve predictive performance by removing redundant or noisy genes that only serve to overfit the model. The authors note, unfortunately, that it is usually very difficult to perform all three simultaneously.

There are two general methods for feature selection (and can be used together). One is using domain knowledge *via* feature engineering and one is utilizing automated approaches. In feature engineering, a domain expert may pick and choose variables from a larger pool that he or she thinks are important prior to more formal model selection. As discussed in Section "Rashomon Effect," this bias can often lead to spurious conclusions when different

research groups pre-select their variables (69). In many genomics applications, often precurated gene ontology data are referenced at some point through a hypothesis-driven approach, either as an initial screen or as part inferring functional relationships after significant genes have been selected. This does introduce a bias toward highly studied gene functions or pathways and a bias against undiscovered gene function, which reinforces the importance of hypothesis-generating studies (see The Importance of Fishing).

Below, we discuss automated approaches for feature selection. The first two are general approaches that are either pre-processing features through a method independent or dependent of the final predictive model. A third approach is to transform the existing features to create new synthetic features (91).

2.5.1. Pre-Processing Variables Independent of the Prediction Model

Filtering (or ranking) variables is the least computationally intensive method for feature selection. This method involves selecting features prior to training a model and is thus independent of the model choice. A common method is to perform univariate correlation testing (for continuous variables) or receiver operating curve analysis (for categorical variables) and then only choosing the top-ranking variables. While efficient in that the processing time scales linearly with the number of variables, filtering does not screen out highly correlated features—in fact, these will be more likely to be selected together. However, Guyon and Elisseeff did show that presumably redundant variables can decrease noise and consequently improve classification (91). Statistically, filtering variables is robust against overfitting as it aims to reduce variance by introducing bias (92). Univariate filtering methods do not consider interactions between features, and thus is unable to assist in determining what variable combination is optimal. In GWAS, statistical tests for univariate significance are an example of variable filtering and thus are unable to account for multi-locus interactions (93). This weakness is magnified when a variable that is uninformative by itself gains value when combined with another variable, as is proposed in epistasis; in this case, filtering would remove the univariately useless variable before it can be tested in combination with another variable. To address this weakness, filter methods such as the ReliefF family take a multivariate and ensemble approach to yield variable rankings (94–96).

2.5.2. Embedding Feature Selection With the Prediction Model

Combining feature selection with the model establishes a dependence that can be used to address issues with multicollinearity and feature interactions. Wrappers combine feature selection with model building but are computationally expensive (97). Various search strategies can be utilized, but often used are greedy search strategies where predictors are either added or removed one-by-one *via* forward selection or backward elimination, respectively. In regularization, feature selection is built into a method's objective function (i.e., the optimization goal) through penalty parameters. These penalty parameters ensure that feature importance (weight) and/or number is incorporated during model training. Common regularization methods include L1-norm or lasso regression (98), L2-norm or ridge regression (99), and combined L1–L2 or elastic networks (100). Regularization methods are of significant interest in applications of ML to genomics due to their ability to decrease the complexity of a polygenic problem and improve probability of replication (90). A relatively novel method developed for feature selection in very high dimensions is stability selection, which uses subsampling along with a selection algorithm to select out important features (101).

2.5.3. Feature Construction and Transformation

Instead of working directly with the given features, features can be manipulated to reconstruct the data in a better way or to improve predictive performance. There are many methods that can perform feature construction with different levels of complexity. Clustering is a classic and simple method for feature construction that replaces observed features by fewer features called cluster centroids (102). Principal component analysis (PCA) provides a method related to eigenvector analysis to create synthetic features which can explain the majority of the information in the data; for example, PCA can decrease type I error by uncovering linkage disequilibrium (LD) patterns in genome-wide analyses due to ancestry (103, 104). Kernel-based methods such as SVMs also make use of feature transformation into higher dimensions and will be discussed in a later Section "Support Vector Machines." Neural networks are another popular ML method that specializes in constructing features within the hidden layers after being initialized with observed features. In the last few years, neural networks have become extremely popular in the form of deep learning, which is discussed below.

2.6. Deep Learning

"Deep learning" describes a class of neural networks that has exploded in popularity in the recent years—particularly in the fields of computer vision (105) and natural language processing (106)—as larger training data sets have become available and computational processing resources have become more accessible and affordable (107). Deep learning is distinguished from earlier neural network methods by its complexity: whereas a "shallow" neural network may have only a few hidden layers, deep learning networks may have dozens (108) to hundreds of layers (109) where unsupervised, hierarchical feature transformation can occur. In popular science, deep learning is the artificial intelligence powering IBM Watson (110) and autonomous driving vehicles. Within medical research, there have been several high-profile deep learning publications claiming expert-level diagnostic performance (111–114). A related domain is radiomics, which seeks to use ML and statistical methods to extract informative imaging features or "phenotypes" in medical imaging (115–117) with a significant focus in oncology imaging (118–121). Deep learning is in an early stage within genomics, but has been used for discovery of sites for regulation or splicing (122, 123), variant calling (124), and prediction of variant functions (125). For further reading on deep learning, we recommend Lecun et al.'s excellent review (107).

3. ML IN GENOMICS

Genomics presents a challenging problem for ML as most methods were not originally developed for GWAS, and thus improving implementations remain a topic of ongoing research (126). The quantity of genomics data recommended for finding significant SNPs is more akin to that seen in image processing, where there could be tens of millions of voxels in a typical computed tomography scan. Given the imbalance of features compared with samples (the "$p \gg n$" problem), there is a challenge in creating predictive models that do not overfit. As discussed in Section "Current ML Approaches to Radiogenomics," different ML methods have been used to address different concerns in genomics and radiogenomics.

In this section, we will review some of the intuition and principles behind genomics methods to better understand how to improve and apply them to future problems.

3.1. Multiple Hypothesis Correction

Hypothesis testing is a principle based on statistical inference. In GWAS, however, one is not just testing a single hypothesis, but millions. As such, by random chance, it is a virtual guarantee that some of the associations will appear to be statistically significant if there is no correction to the pre-specified significance level α (127). How to correct for multiple hypothesis comparisons is an area of significant interest in GWAS and there are many techniques to do so (128). These methods generally aim to control the number of type I errors and include family-wise error rate (FWER)—the probability of at least one type I error—and false discovery rate—the expected proportion of false discoveries (129). Controlling FDR has greater power than FWER at the risk of increased type II error (130). One common FWER correction method is Bonferroni correction, which would work reasonably well for independent tests, but is an overly strict (i.e., conservative) bound for GWAS due to the prevalence of LD across the genome. LD causes adjacent regions of the genome to be inherited together, and thus Bonferroni will overcorrect due to non-independence among SNPs within LD blocks. For rare variants which are not thought to be in LD, Bonferroni correction would be an appropriate correction.

In ML, poor correction for multiple testing is related to p-hacking or data dredging, which is to continuously run iterations of this method until it fits a pre-conceived notion or hypothesis (131) (see Lessons From Statistics).

3.2. The Case of Missing Heritability

As sample sizes have increased since the first GWAS in 2005, more and more robust associations with loci have been discovered in genomics (132). This has also been reflected in radiogenomics as larger sample sizes have been possible through the RGC (see Genomic Basis for Radiotherapy Response). However, the discovered associations are still relatively few and insufficient to explain the range of observed phenotypes, creating the so-called "case of missing heritability" (133). Response of both normal and tumor tissue has certainly shown itself to be a complex, polygenic trait (29, 30, 54). The cause of this missing heritability is thought to arise from several sources, including common variants of low effect size, rare variants, epistasis, and environmental factors. One clear solution already underway is to genotype more samples and to use meta-analysis methods to combine results across studies (134). However, there are limits to this approach. For one, rare variants [minor allele frequency (MAF) < 0.0005] with smaller effect sizes (odds ratios ~1.2) will require between 1 and 10 million samples for detection using standard GWAS techniques (132). Another issue is that epistatic interactions among common variants have not been able to be reliably replicated (77). ML provides a complementary approach for finding patterns in noisy, complex data and detecting non-linear interactions.

3.3. Combining ML and Hypothesis Testing

Originally, two-stage GWAS was developed from standard one-stage GWAS to decrease genotyping costs in an era where SNP chips were costlier (135). In this method, all SNP markers are genotyped in a proportion of the samples in stage 1, and a subset of the SNPs would then be selected for follow-up in stage 2 on the remaining samples. This method does not decrease type I or II error, however (136). Performing a joint analysis where the test statistics in stage 2 were conditional on stage 1 had superior results than assuming independence between the two stages (i.e., a replication study), but power is unable to exceed that of one-stage GWAS (137). Instead of two-stage GWAS, a promising alternative is to use two-stage models combining ML and statistical hypothesis testing, aiming to combine the strengths of separate methodologies (see Statistical Inference vs. ML). These combined models can increase power and uncover epistatic interactions (138).

3.3.1. Learning Curves and Power

In principle, combining ML and hypothesis testing works because, by design with setting a pre-determined alpha level and power, statistical inference does not benefit from larger datasets once a result has met statistical significance. Indeed, larger datasets can result in detection of statistically significant associations of decreasing effect size and potentially decreasing clinical relevance. This limitation does not apply to ML, which can asymptotically use more data to improve predictive performance. Many ML methods are characterized by a learning rate obeying an inverse power law with respect to sample size (80, 139, 140). This behavior suggests that ML offers a complementary approach to statistical methods by continuing to learn for each additional sample. With increasing sample sizes and meta-analyses, one can imagine a scenario where one is well in the "plateau" portion of the power curve and can afford samples to be used in the ML method (**Figure 3**).

3.3.2. Using ML to Detect Epistasis

Epistasis, which includes interactions between SNPs, is not well accounted for in standard GWAS. Epistatic interactions are recognized as a cause of non-linear effects and may help elucidate functional mechanisms as well (141). Biological interpretations of epistasis have been difficult with little correlation between statistical interaction and physical interaction (i.e., protein–protein binding) or other biologic interactions (142). Regardless

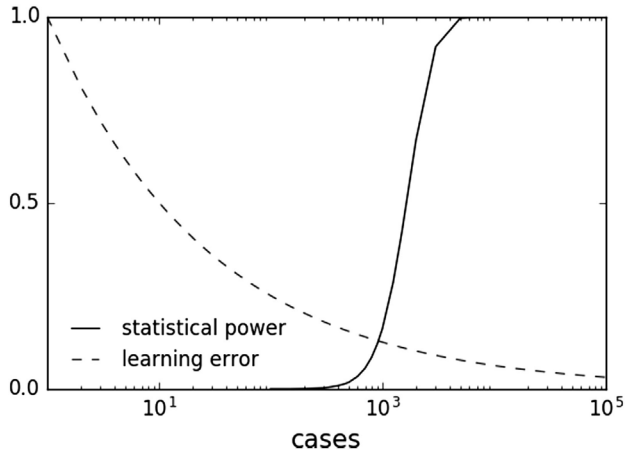

FIGURE 3 | Sample plots of statistical power and learning curve error. Statistical power graph derived using Genomic Association Studies power calculator (137). Learning curve assuming an inverse power law common to multiple machine learning methods (80, 139, 140).

of whether protein products are physically interacting with other proteins or environment, the statistical interaction suggests that there is dependence at some level for a specific disease (141).

Given the exponentially increased search space for SNP interactions, there is a high concern for false positives (see The Curse of Dimensionality). This concern is magnified when SNPs are in LD. A filtering method is often used to decrease the search space for only the most promising interactions (see Pre-Processing Variables Independent of the Prediction Model). Exhaustive searches for pairwise interactions are also now becoming possible, aided by the massive advances in parallel processing throughout offered by graphical processing units (143, 144).

Due to technical limitations in accounting for non-linear effects and multiple hypothesis correction in an exhaustive search, interaction studies have typically focused on SNPs with weak marginal effects (77). Unfortunately, many of the studies in non-cancer diseases have not been successful (145, 146). One postulate is that pairwise SNPs are unlikely to have large interaction effects. However, as sample sizes and SNP density improve (to better tag causal variants while avoiding spurious interactions due to LD), then ML methods that incorporate SNP interactions with low or no marginal/main effects may begin to uncover replicable interaction effects (138, 147–149).

Two-stage methods are a promising approach that combines the strength of fast, approximate interaction tests with a subsequent thorough model (77). Such methods take advantage of the strength of statistical tests for detecting polygenic low signal, linear interactions with the ability of ML to train cross-validated models of non-linear interactions (150, 151). Regularization within two-stage methods is an area of interest (90). Wu et al. adapted lasso to LR for use in dichotomous traits in GWAS (152). Wasserman and Roeder developed a similar procedure called "screen and clean" that also controls for type I error by combining lasso linear regression, cross-validated model selection, and hypothesis testing (153). Like traditional two-stage GWAS, the data are split between the stages. Wu et al. adopted this model to model interaction effects in addition to main effects (154).

As further discussed in Section "Random Forest," ensemble tree-based methods are very popular for detection of epistatic interactions (148, 155, 156). While it is difficult to assess statistical significance in ensemble black box techniques, permutation re-sampling methods can be used to determine a null distribution and associated *p*-values (80, 138, 141) (see CV Relationship With Statistical Inference). Other popular methods for interaction that have continued to receive updates include a cross-validated dimensionality reduction method called multifactor dimensionality reduction (157) and a Markov Chain Monte Carlo sampling method to maximize posterior probability called Bayesian Epistasis Association Mapping (158).

3.3.3. Using ML to Increase Power

Overfitting and false discoveries (type I errors) represent similar concepts in ML and statistical inference, respectively, in that both falsely ascribe importance. Like the bias-variance tradeoff, statistical inference seeks to balance type I and type II errors. As each hypothesis test represents an additional penalty to genome-wide significance, one way to decrease type II error is to decrease the number of hypothesis tests. While decreasing testable hypotheses may appear to decrease power, Skol et al. demonstrate that being more stringent in selecting SNPs in stage 1 may paradoxically increase power as the multiple testing penalty is subsequently reduced in stage 2 (137).

Combination of ML and statistical methods can simultaneously be designed to detect epistasis and increase power (138). In "screen and clean" (see Using ML to Detect Epistasis), Wasserman and Roeder perform L1-regularization in the "clean" phases to improve power in the "screen" phases. Meinshausen et al. extend the method by Wasserman and Roeder by performing multiple random splits (instead of one static split) to decrease false positives and increase power (159). Mieth et al. similarly combined SVM with hypothesis testing (160), but instead of splitting, they re-sample data using an FWER correction (161). While re-sampling for feature selection and parameter tuning may bias toward more optimistic results (see Freedman's Paradox), Mieth et al. report higher power compared with Meinshausen and Wasserman and Roeder, with 80% of the discovered SNPs validated by prior studies. Nguyen et al. took a similar approach except with RF instead of SVM (162).

Combined ML and statistical methods can either have the ML stage first or second. When ML is used first, it usually acts as a feature selection filter to reduce the multiple hypothesis penalty and increase power for hypothesis testing in the second stage. When the ML step is second, it acts to validate candidate SNPs that passed the first stage filter. The order of ML and hypothesis testing may not affect power. Mieth et al. report similar results compared with Roshan et al. (163), who performed chi-square testing followed by RF or SVM [supplement in Ref. (160)]. Similarly, Shi et al. proposed single SNP hypothesis testing followed by lasso regression, which was the reverse order of Wasserman and Roeder (164).

Oh et al. used a multi-stage approach to uncover novel SNPs and improve prostate radiotherapy toxicity prediction (165, 166).

The first step is to create latent (indirectly observed) variables through PCA. These "pre-conditioned" variables are fit using LR to the original outcomes. This serves to create "pre-conditioned" outcomes that are continuous in nature and provides estimate of radiotoxicity probability. These pre-conditioned outcomes are then modeled using RF regression and validated on holdouts of the original samples.

4. CURRENT ML APPROACHES TO RADIOGENOMICS

Machine learning models are particularly attractive when dealing with genetic information, as they can consider SNP–SNP interactions, which are suspected to be important, but are often missed by classical association tests because their marginal effects are too small to pass stringent genome-wide significance thresholds.

However, ML models also come with constitutional pitfalls, namely, increased computational complexity and risk for overfitting, which must be acknowledged and understood to avoid reporting impractical models or over-optimistic results.

Current use of ML techniques in radiogenomics usually follows the top-down approach, where radiotherapy outcomes are modeled through complex statistical analysis, without considering *a priori* knowledge of interactions of radiation with tissue and biological systems. In this field, supervised learning is widely preferred, i.e., models aim at constructing a genotype–phenotype relationship by learning such genetic patterns from a labeled set of training examples. Supervised learning can provide phenotypic predictions in new cases with similar genetic background. Nevertheless, an unsupervised approach (e.g., PCA or clustering) is sometimes used to reduce the dimensionality of datasets, extract a subset of relevant features, or construct features to be later included in the chosen learning method. Feature selection is of extreme importance (see Feature Selection), as it leads to the reduction of the dimensionality of the genetic search space, excluding correlated variants without independent contribution to the classification, and helping the translation of the model to the clinical setting.

Even if most ML techniques can act both as regression and classification methods, the classification or discriminative aspect has been most investigated in recent years, with main interest in separation between patients with/without the selected study outcome (e.g., presence/absence of radiotherapy-induced toxicity, tumor control/failure, and presence/absence of distant metastasis).

There is also increasing interest in overcoming the "black box" characteristics of some ML methods, favoring use of techniques that allow ready interpretation of their output (see Occam Dilemma), making apparent to the final user the relationships between variables and the size and directionality of their effect, i.e., if the variables are increasing or decreasing the probability of the outcome and the magnitude of their impact.

In this frame, RF, SVMs, and Bayesian networks (BNs) received great attention and they constitute the main topic of this section (**Table 1**). The presented ML algorithms can accommodate GWAS-level data. When considering the emerging sequencing domain (e.g., whole-exome and genome profiling), new technical challenges are posed that might be addressed by new algorithmic advances or by parallelization and cloud technologies for distributed memory and high-performance computing.

4.1. Random Forest

Random forest is a regression and classification method based on an ensemble of decision trees (172). The ensemble approach averages the predicted values from individual trees to make a final prediction, thus sacrificing the interpretability of standard decision trees for increased prediction accuracy (74). Each tree is trained on bootstrapped training samples (i.e., sampling with replacement), while a random subset of features is used at each node split. When applied to a problem of predicting a disease state using SNPs, for example, each tree in the forest grows with a set of rules to divide the training samples based on discrete values of the genotypes (e.g., homozygous vs. heterozygous). Here, we list the characteristics of RF that make it an attractive choice for GWAS, both for outcome prediction and hypothesis generation.

4.1.1. Robustness at High-Dimensional Data

Given high-dimensional data, training predictive models likely faces risk of overfitting. The ensemble approach utilized by RF mitigates this risk by reducing model variance due to aggregation of trees with low correlation. Examples of studies emphasizing predictive performance of RF include work by Cosgun et al. (174), Nguyen et al. (162), Oh et al. (165) (SNP based), Wu et al. (175), Díaz-Uriarte and Alvarez de Andrés (176), and Boulesteix et al. (177) (microarray based). While RF was initially thought not to overfit based on datasets from the UCI ML repository (65), this was ultimately found to be incorrect when noisier datasets were

TABLE 1 | Three representative machine learning methods with select pre-processing tips and tuning methods for complexity control.

Method	Pre-process	Complexity control	Reference
Support vector machine (SVM)	– Encode features as binary – Normalize to uniform distribution – Imputation for balancing data	– Recursive feature elimination for linear SVM – Soft margin width (C-parameter) – Kernel hyperparameters	(76, 160)
Bayesian networks	– Feature discretization – Variable selection to reduce graph search space – Imputation not necessary when using expectation maximization	– Constraints to a graph search space based on prior knowledge – Graph scoring functions that penalize complexity	(167–171)
Random forest	– No discretization or normalization necessary – Imputation required	– Number of features to sample at each node split (mtry) – Minimum number of samples in a terminal node	(172, 173)

introduced (178). When training RF models, some parameters need to be optimized, which can affect predictive power. Among those, the number of variables that are randomly selected from the original set of variables at each node split (*mtry)* governs model complexity. Many studies opt for default configurations as originally recommended by Breiman (172) (classification: $\sqrt{p}$, regression: $p/3$ where p: number of predictors), and predictive performance was shown to be stable around these values (176, 179). However, a larger *mtry* is recommended when there are many weak predictors (172), which might be the case for GWAS of complex diseases. Goldstein et al. (173) conducted a search for optimal parameters in GWAS of multiple sclerosis, comprising about 300K SNPs, and recommended $mtry = 0.1$ after initial pruning of the SNPs under high LD.

4.1.2. Biomarker Prioritization

Random forest can provide a variable importance measure (VIM), which quantifies the influence of an individual predictor on the purity of the node split (purity based) or prediction accuracy in unseen samples (permutation based). VIM can be used for selecting a smaller subset of genes or SNPs from GWAS, which can be further used for achieving higher predictive performance or biological validation. Lunetta et al. (180) proposed to use RF VIM for SNP prioritization as an alternative to Fisher's p-value under the presence of SNP–SNP interactions. Nguyen et al. (162) used VIM as a feature selection process for a subsequent RF training to enhance predictive performance. However, reliability of VIM, especially under LD, has been questioned and investigated by simulation studies: Tolosi and Lengauer (181) and Nicodemus et al. (182) suggested that VIM may not correctly measure the importance of a large group of correlated SNPs due to dilution of VIM. Also, Strobl et al. (183) showed potential bias in VIM toward the predictors with more categories; they proposed the conditional inference tree as an alternative where each node split is performed based on a conditional independence test instead of the conventional Gini index (184).

4.1.3. Ability to Account for SNP–SNP Interactions

Epistasis describes the non-linear combination of SNPs (or SNP and environment) that may correlate with a phenotype. Epistasis is thus important for understanding complex diseases (77). By construction, RF can indirectly account for epistasis through successive node splits in a tree where one node split is conditional upon the split from the previous node. Lunetta et al. (180) claimed that RF VIM has a higher power of detecting interacting SNPs than univariate tests. Thus, RF has been used as a screening step to identify much smaller number of SNPs that are more likely to demonstrate epistasis, which can be further tested in a pairwise fashion (150, 151). However, Winham et al. (156) warned that ability of RF VIM to detect interactions might decrease with an increasing number of SNPs and large MAF of SNPs.

4.1.4. Hybrid Methods

Random forest is occasionally used in conjunction with other ML methods. Boulesteix et al. (177) used partial least squares to reduce dimensionality of gene microarray data prior to training a RF classifier. Stephan et al. (185) used RF as a fixed component of a mixed-effect model to handle population structure. Oh et al. (165) introduced a pre-conditioning step prior to RF training where a binary outcome of radiotherapy toxicity was converted to a continuous pre-conditioned target, which helps reduce the noise level that may be present in the outcome measurements (186).

4.2. Support Vector Machines

Support vector machines are usually used to solve the problem of supervised binary classification. In the field of oncologic modeling, SVMs are used to classify new patients into two separate classes (with/without the outcome of interest) based on their characteristics (76). The first step is to find an efficient boundary between patients with/without the outcome in the training set. This boundary is called a "soft margin" and is a function of the known d features of the patients included in the training set. To determine this boundary, non-linear SVMs use a technique called the kernel trick to transform data into a higher dimension, whereby they can then be separated by a d-dimensional surface in a non-linear fashion. Based on these transformations, SVM finds an optimal boundary between the possible outcomes. In technical terms, a linear SVM models the feature space (the space of possible support vectors, which is a finite-dimensional vector space where each dimension represents a feature) and creates a linear partition of the feature space by establishing a hyperplane separating the two possible outcomes. Of note, the created partition is linear in the vector space, but it can use the kernel trick to solve non-linear partition problems in the original space. Based on the characteristics of a new patient, the SVM model places the new subject above or below the separation hyperplane, leading to his/her categorization (with/without the clinical outcome). SVMs maximize the distance between the two outcome classes and allow for a defined number of cases to be on the "wrong side" of the boundary (i.e., a soft margin). Due to this, despite the complexity of the problem, the SVM boundary is only minimally influenced by outliers that are difficult to separate.

Support vector machines are a non-probabilistic classifier: the characteristics of the new patients fully control their location in the feature space, without involvement of stochastic elements. If a probabilistic interpretation for group classification is needed, the measure of the distance between the new patient and the decision boundary can be suggested as a potential metric to measure the effectiveness of the classification (187).

4.2.1. Robustness in High-Dimensional Data and Possibility to Handle for Variable Interaction

Support vector machines are particularly suited to model datasets including genomic information, as they are tailored to predict the target outcome (the phenotype) from high-dimensional data (the genotype) with a possible complex and unknown correlation structure by means of adaptable non-linear classification boundaries. The framework of SVMs implicitly includes higher-order interactions between variables without having to predefine what they are. Examples of studies highlighting good performance of SVMs in this area are (188–190).

The main pitfall presents when the number of variables for each patient exceeds the number of patients in the training dataset. For this reason, in such case, the combination of SVMs

with techniques aimed at reduction of the number of features is suggested.

Support vector machines can be used to approach analysis of GWAS data even in combination steps. Mieth et al. (160) proposed a two-step SVM procedure with SVMs first adopted for testing SNPs by taking their correlation structure into account and for determining a subset of relevant candidate SNPs (see Combining ML and Hypothesis Testing). Subsequently, statistical hypothesis testing is performed with an adequate threshold correction. As complexity reduction is performed prior to hypothesis testing, the strict multiple correction threshold can thus be relaxed.

4.2.2. Tuning Parameters

Considering practical challenges in SVM modeling, a key issue is tuning the parameters identifying the separation hyperplane and determining how many support vectors must be used for classification. There are also kernel-specific parameters to tune. Grid search is traditionally used to find the best set, with choice of initial conditions and search strategy highly influencing the quality of the result (191, 192).

4.2.3. Unbalanced Datasets

Attention must also be paid when SVMs are applied to unbalanced data, i.e., one outcome class contains considerably more cases than the other. This scenario is common in radiotherapy modeling where toxicity and local failure rates can be low. Unbalanced datasets present a challenge when training every type of classifier, but particularly is true for maximum-margin classifiers such as SVM. A satisfactory choice for having a high-accuracy classifier on a very imbalanced dataset could be to classify every patient as belonging to the majority class. Nevertheless, such a classifier is not very useful. The central issue is that, in such a case, the standard notion of accuracy is a bad measure of the success of a classifier, and a balanced success rate should be used in training the model, which assigns different costs for misclassification in each class (170, 193, 194). These methods can include showing a full confusion matrix; reporting F1-score and positive/negative predictive values, which incorporate relative imbalances (195–197); or synthetic balancing through undersampling and/or oversampling (198).

4.2.4. Interpretation of SVMs

Interpreting SVM models is far from obvious. Consequently, work is being done in providing methods to visualize SMV results as nomograms to support interpretability (199, 200).

The absence of a direct probabilistic interpretation also makes SVM inference difficult, with the aforementioned work by Platt being one solution (187).

4.3. Bayesian Networks

Bayesian network is a graphical method to model joint probabilistic relationships among a set of random variables, meaning that the variables vary in some random or unexplained manner (201). Based on the analysis of input data or from expert opinion, the BN assigns probability factors to the various results. Once trained on a suitable dataset, the BN can be used to make predictions on new data not included in the training dataset.

A key feature of BN is graphical representation of the relationships *via* a directed acyclic graph (DAG). Although visualizing the structure of a BN is optional, it is a helpful way to understand the model. A DAG is made up of *nodes* (representing variables) and directed *links* between them, i.e., links originate from a parent variable and are pointed to child variables without backwards looping or two-way interactions. Parent variables influence the probability of child variables and the probability of each random variable is established to be conditional upon its parent variable(s). In this way, the DAG encodes the presence and direction of influence between variables, which makes BN attractive for users needing intuitive interpretation of results (169) (see Occam Dilemma). This directionality of links is important as it defines a unique representation for the multiplicative partitioning of the joint probability: the absence of an edge between two nodes indicates conditional independence of involved variables.

4.3.1. Interpretation of BNs

Bayesian networks can integrate different data types into analysis. Despite accounting for high-order variable interactions (e.g., genetic environment), BNs maintain high interpretability *via* graphical outputs. As an example, **Figure 4** demonstrates a possible BN for prediction of radiotherapy-induced rectal bleeding following different clinical, genetic, and treatment-related variables.

4.3.2. Using Knowledge and Data in a Synergistic Way

A DAG can be built starting from previous knowledge, or completely trained on available data. For example, BN was used to incorporate expert knowledge along with experimental assay data

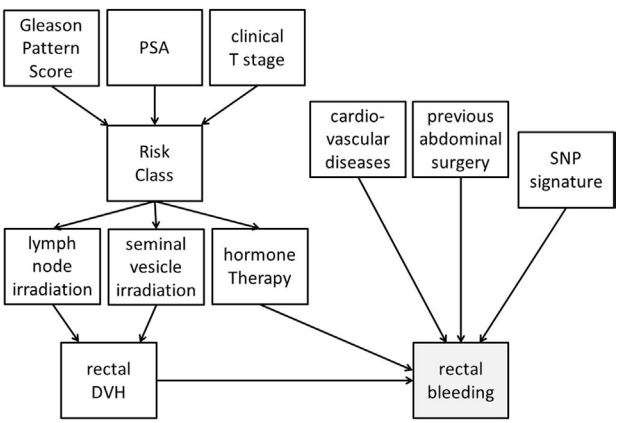

FIGURE 4 | Possible representation of a Bayesian network directed acyclic graph for predicting late rectal bleeding after radiotherapy for prostate cancer. The network includes tumor-related characteristics (PSA, Gleason pattern score, and clinical T stage) which determine risk class and consequently radiotherapy targets (irradiation of pelvic lymph nodes and of seminal vesicles) and use of concomitant hormone therapy. Treatment variables influence the dosimetry of organs at risk [rectal dose–volume histogram (DVH)], and this has a causal effect on late rectal bleeding probability. Clinical (presence of a previous abdominal surgery and of cardiovascular diseases) and genetic [single-nucleotide polymorphism (SNP) signature] variables with (causal) associations with rectal bleeding are also included in the DAG.

to assign functional labels to yeast genes (202). The optimized DAG is the one which maximizes a predefined scoring function over all possible DAG configurations. When multiple DAGs score at the same level, an approach embracing an ensemble of models can be followed (169).

4.3.3. Robustness at High-Dimensional Data

Since the number of possible DAGs grows super-exponentially with the number of available features, it is unrealistic to comprehensively search for the highest-scoring DAG over all graph possibilities. This is especially true when considering high-dimensionality problems encountered in GWAS. Various approaches could be suggested to confront the burden (169, 170):

(a) Use a causality prior that considers the already available knowledge to impose restrictions on the presence/direction of links between nodes to reduce the search space.
(b) Structure features into systems of different hierarchical levels with connections established by combining data and prior knowledge.
(c) Reduce input dimension by appropriate variable selection techniques with the aim of removing highly correlated features.
(d) Use of graph scoring functions that penalize complex graph structures, such as Bayesian information criteria (167).

An interesting approach is also the use of a forest of hierarchical latent class models (171) to reduce the dimension of the data to be further submitted to BN to discover genetic factors potentially involved in oncologic outcomes. Latent variables are thought to capture the information coming from a combination of SNP, genetic, and molecular markers. Latent variables can also be clustered into groups and, if relevant, such groups can be subsequently incorporated into additional latent variables. This process can be repeated to produce a hierarchical structure (a forest of latent variables) and BN analyses can be primarily completed on latent variables coupled to a largely reduced number of clinical and dosimetric features.

4.3.4. Handling Missing Values

The probabilistic approach of BNs makes them suitable to efficiently handle missing values, without removal of cases or imputation. A BN can be trained even using non-complete cases and it can be queried even if a full observation of relevant features is not available. This is an advantage in clinical oncology where missing data are the norm and not the exception.

Bayesian networks were successfully applied in many oncologic/radiotherapy studies, including modeling of radiation-induced toxicity, tumor control after radiotherapy, and cancer diagnosis (169, 170, 203–207).

5. IMPROVING ML INTEGRATION IN RADIOGENOMICS

Machine learning holds significant promise for advancing radiogenomics knowledge through uncovering epistatic interactions and increasing power. In this section, we will discuss general lessons learned and potential barriers.

5.1. Lessons From Statistics

For ML models to focus on predictive performance alone while not taking lessons from statistical theory would be a mistake. Statistical genetics learned through many iterations that it is necessary to take into account multiple hypothesis testing to decrease type I error (127). While ML models are often framed to be hypothesis-free, they can fall into a trap of cherry picking results that show good performance, which may end up being spurious. This practice of trawling for results that appear statistically significant has been called data dredging or p-hacking and has been cautioned against by the American Statistical Association (131). However, this practice can occur surreptitiously, such as when a pharmaceutical drug is tested in many highly correlated trials (i.e., asking similar questions) over many years, but without correcting for multiple testing. This phenomenon is particularly common in oncology where there is vested interest to find an application for a "blockbuster" therapeutic (208, 209). One solution for this is to create drug development portfolios to apply meta-analysis principles to drug trials instead of considering them as individuals (210). A similar approach could be used in radiogenomics to avoid publication bias and report negative results.

Notably, in their same report, the American Statistical Society emphasizes a distinction between statistical significance and clinical significance. Whether a *p*-value does or does not meet an α cutoff does not preclude it from being validated. ML provides an excellent tool for validation when used in the two-step models.

5.2. Reusable Hold-Out Set

Due to the nature of model building, it is often desirable to repeatedly refine one's model due to suboptimal performance on the independent "holdout" set. Unfortunately, as discussed earlier (see Common Errors in CV), re-testing presents a significant problem as the refined model is now biased by newly obtained knowledge. For example, one might manually curate variables or alter hyperparameters to try to improve test set performance repeatedly, leading to overfitting on a true external dataset. However, reserving multiple test sets is not practical in most projects. One intriguing solution arose from university–industry collaborations with technology companies such as IBM, Microsoft, Google, and Samsung (211). These companies are interested in differential privacy, which is the concept of preserving the privacy of an individual while still collecting aggregate group statistics (212). This is not a trivial problem as knowledge about an aggregate sample over time can precisely identify supposedly "anonymous" individuals. For example, measuring the mean of a sample before and after removing one data point would allow one to precisely determine the value of that one data point if one knew the sample size. A prominent example in 2008 involved de-anonymizing publicly released Netflix data using another website (the Internet Movie Database) to ascertain apparent political affiliations and other potentially sensitive details (213). Differential privacy concepts are directly related to the necessity of maintaining independence—in essence, the "anonymity"—of the holdout set. These concepts have been adapted to a reusable holdout, whereby the holdout can be resampled many times through a separate algorithm (211, 214, 215). The number of

times that the holdout can be reused grows roughly with the square of its size, thus potentially providing near-unrestricted access for large datasets such as GWAS.

5.3. Incorporate Clinical Variables

Many complex disease phenotypes are likely confounded by environmental effects. When genetic and environmental determinants are combined, there is increased accuracy in heritability prediction (216). This contribution from an environmental, non-genetic source suggests that multi-domain models incorporating both genetic and clinical factors should create a superior predictor compared with genetic predictors alone. Current radiotherapy prediction models focus on clinical and dosimetric variables but do not incorporate genetic factors (217). Both the ASTRO and the European Society of Radiation Oncology recognize a need for improved radiation toxicity models—including through ML (218)—and have pushed for utilization of big data toward "precision" radiation oncology (219, 220).

5.4. Replication and Regulatory Concerns

When applying ML to radiogenomics for eventual human applications, one must also consider practical concerns about the current regulatory environment. In the mid-late 2000s, a wave of multi-biomarker laboratory-developed tests (LDTs) in oncology emerged that made several bold, highly publicized promises. Some were met (see Precision Medicine and Multigene Panels) but many ultimately went unfulfilled. These included two proteomics-based diagnostic tests for ovarian cancer. OvaCheck (221, 222) was debunked due to data artifacts (223) and batch effects (224). OvaSure (225, 226) was pulled from market in 4 months after FDA intervention due to concerns for inadequate validation (227). Both tests reported overly optimistic positive predictive values due to being trained on unrealistic data of approximately 50% cancer positivity, whereas true ovarian cancer incidence is closer to 1 per 2,500 post-menopausal women (195–197, 227) (see Unbalanced Datasets). Certainly, the most high-profile and drawn-out case (85) involved lung cancer genomics-based chemotherapy response prediction that was pre-maturely rushed to clinical trial (228–230). Investigations into these and other controversies surrounding poor understanding of statistics and independent validation in biomarker studies (see Rashomon Effect) led to an extensive report by the Institute of Medicine which suggested corrective measures (84). Controversy continues regarding whether and how the FDA should regulate LDTs while still promoting innovation (231). One potential direction is pre-certifying laboratories instead of individual LDTs. Regardless, understanding modeling principles in a scientific environment increasingly reliant on big data analysis is necessary to avoid repeating the same mistakes of a decade ago.

5.5. Promoting Research

An executive summary from the ASTRO Cancer Biology/Radiation Biology Task Force (232) and a report from the ASTRO/AAPM/NCI 2016 precision medicine symposium (6) both recognized the large relative disparity between the utilization of therapeutic radiation (between 50 and 66% of cancers) and its investigative research effort. In the US, there are approximately 5,000 radiation oncologists and 15,000 medical oncologists, but a 2013 review of US National Institutes of Health (NIH) funding in radiation oncology found that <50% of all accredited departments had an active research program with at least 1 NIH grant, which is at odds with radiation oncology attracting the highest percentage of MD/PhD residents for a number of years (233). Only 3% of successfully awarded grants by the NIH Radiation Therapeutics and Biology study section are for biomarkers or radiogenomics (232). These numbers suggest that radiogenomics research continues to be underfunded. While the field moves toward improved support of young investigators through opportunities like the Holman Pathway (234, 235) and more is discovered in radiobiology and radiogenomics, there will also be a need to support methods development to ensure that radiation oncology does not lag behind in the era of precision medicine.

6. CONCLUSION

Oncology is a field enriched by multidisciplinary study. Like cancer, genetics has eluded a complete understanding due to its surprising level of complexity. The focus on ML in the technology industry is quickly moving into medicine, with a prime example being IBM Watson's ability to understand game show questions becoming adapted for tumor board recommendations (114). These translational research efforts are not easy and require teamwork from stakeholders of varying backgrounds to avoid repeating mistakes made in one field in another field. In a radiogenomics era, radiation oncology will require multidisciplinary integration of not just radiation biologists, physicists, and oncologists but also insight from computational biologists, statistical geneticists, and ML researchers to best treat patients using precision oncology.

AUTHOR CONTRIBUTIONS

JK wrote the manuscript except section IV, which was written by TR, SL, and JO. SLK, JS, RS, SYK, and BR rewrote portions and made edits. All the authors approved the manuscript.

FUNDING

RS is supported by U.S. National Institutes of Health award R21CA216452 and Pennsylvania Department of Health Grant GBMF4554 #4100070287. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions. TR is supported by Associazione Italiana Ricerca sul Cancro (AIRC-IG16087). JO and SL are supported by National Institutes of Health/National Cancer Institute Cancer Center Support Grant (Grant number P30 CA008748). BR is supported by NIH (1R01CA134444, HHSN261201500043C, and HHSN261201700033C), the American Cancer Society (RSGT-05-200-01-CCE), and the Department of Defense Prostate Cancer Research Program (PC074201 and PC140371). SLK is supported by NIH/NCI 1K07CA187546 SYK was supported by Pennsylvania Department of Health Grant GBMF4554 #4100070287 and NSF CAREER Award No. MCB-1149885.

REFERENCES

1. Hall EJ, Giaccia AJ. *Radiobiology for the Radiologist*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins (2012).
2. Mould RF. Pierre curie, 1859–1906. *Curr Oncol* (2007) 14(2):74–82. doi:10.3747/co.2007.110
3. Grantzau T, Overgaard J. Risk of second non-breast cancer after radiotherapy for breast cancer: a systematic review and meta-analysis of 762,468 patients. *Radiother Oncol* (2015) 114(1):56–65. doi:10.1016/j.radonc.2014.10.004
4. Hudson MM, Poquette CA, Lee J, Greenwald CA, Shah A, Luo X, et al. Increased mortality after successful treatment for Hodgkin's disease. *J Clin Oncol* (1998) 16(11):3592–600. doi:10.1200/JCO.1998.16.11.3592
5. Scaife JE, Barnett GC, Noble DJ, Jena R, Thomas SJ, West CM, et al. Exploiting biological and physical determinants of radiotherapy toxicity to individualize treatment. *Br J Radiol* (2015) 88(1051):20150172. doi:10.1259/bjr.20150172
6. Hall WA, Bergom C, Thompson RF, Baschnagel AM, Vijayakumar S, Willers H, et al. Precision oncology and genomically guided radiation therapy: a report from the American Society for Radiation Oncology/American Association of Physicists in Medicine/National Cancer Institute Precision Medicine Conference. *Int J Radiat Oncol Biol Phys* (2018) 101(2):274–84. doi:10.1016/j.ijrobp.2017.05.044.
7. Baumann M, Krause M, Overgaard J, Debus J, Bentzen SM, Daartz J, et al. Radiation oncology in the era of precision medicine. *Nat Rev Cancer* (2016) 16(4):234–49. doi:10.1038/nrc.2016.18
8. Kachnic LA, Winter K, Myerson RJ, Goodyear MD, Willins J, Esthappan J, et al. RTOG 0529: a phase 2 evaluation of dose-painted intensity modulated radiation therapy in combination with 5-fluorouracil and mitomycin-C for the reduction of acute morbidity in carcinoma of the anal canal. *Int J Radiat Oncol Biol Phys* (2013) 86(1):27–33. doi:10.1016/j.ijrobp.2012.09.023
9. Nutting CM, Morden JP, Harrington KJ, Urbano TG, Bhide SA, Clark C, et al. Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial. *Lancet Oncol* (2011) 12(2):127–36. doi:10.1016/S1470-2045(10)70290-4
10. Chun SG, Hu C, Choy H, Komaki RU, Timmerman RD, Schild SE, et al. Impact of intensity-modulated radiation therapy technique for locally advanced non-small-cell lung cancer: a secondary analysis of the NRG oncology RTOG 0617 randomized clinical trial. *J Clin Oncol* (2017) 35(1):56–62. doi:10.1200/JCO.2016.69.1378
11. Sheets NC, Goldin GH, Meyer AM, Wu Y, Chang Y, Sturmer T, et al. Intensity-modulated radiation therapy, proton therapy, or conformal radiation therapy and morbidity and disease control in localized prostate cancer. *JAMA* (2012) 307(15):1611–20. doi:10.1001/jama.2012.460
12. Folkert MR, Singer S, Brennan MF, Kuk D, Qin LX, Kobayashi WK, et al. Comparison of local recurrence with conventional and intensity-modulated radiation therapy for primary soft-tissue sarcomas of the extremity. *J Clin Oncol* (2014) 32(29):3236–41. doi:10.1200/JCO.2013.53.9452
13. Wang D, Zhang Q, Eisenberg BL, Kane JM, Li XA, Lucas D, et al. Significant reduction of late toxicities in patients with extremity sarcoma treated with image-guided radiation therapy to a reduced target volume: results of radiation Therapy Oncology Group RTOG-0630 trial. *J Clin Oncol* (2015) 33(20):2231–8. doi:10.1200/JCO.2014.58.5828
14. Paumier A, Ghalibafian M, Gilmore J, Beaudre A, Blanchard P, el Nemr M, et al. Dosimetric benefits of intensity-modulated radiotherapy combined with the deep-inspiration breath-hold technique in patients with mediastinal Hodgkin's lymphoma. *Int J Radiat Oncol Biol Phys* (2012) 82(4):1522–7. doi:10.1016/j.ijrobp.2011.05.015
15. Formenti SC, Gidea-Addeo D, Goldberg JD, Roses DF, Guth A, Rosenstein BS, et al. Phase I-II trial of prone accelerated intensity modulated radiation therapy to the breast to optimally spare normal tissue. *J Clin Oncol* (2007) 25(16):2236–42. doi:10.1200/JCO.2006.09.1041
16. Horiot JC, Le Fur R, N'Guyen T, Chenal C, Schraub S, Alfonsi S, et al. Hyperfractionation versus conventional fractionation in oropharyngeal carcinoma: final analysis of a randomized trial of the EORTC cooperative group of radiotherapy. *Radiother Oncol* (1992) 25(4):231–41. doi:10.1016/0167-8140(92)90242-M
17. Turrisi AT III, Kim K, Blum R, Sause WT, Livingston RB, Komaki R, et al. Twice-daily compared with once-daily thoracic radiotherapy in limited small-cell lung cancer treated concurrently with cisplatin and etoposide. *N Engl J Med* (1999) 340(4):265–71. doi:10.1056/NEJM199901283400403
18. Horiot JC, Bontemps P, van den Bogaert W, Le Fur R, van den Weijngaert D, Bolla M, et al. Accelerated fractionation (AF) compared to conventional fractionation (CF) improves loco-regional control in the radiotherapy of advanced head and neck cancers: results of the EORTC 22851 randomized trial. *Radiother Oncol* (1997) 44(2):111–21. doi:10.1016/S0167-8140(97)00079-0
19. Overgaard J, Hansen HS, Specht L, Overgaard M, Grau C, Andersen E, et al. Five compared with six fractions per week of conventional radiotherapy of squamous-cell carcinoma of head and neck: DAHANCA 6 and 7 randomised controlled trial. *Lancet* (2003) 362(9388):933–40. doi:10.1016/S0140-6736(03)14361-9
20. Schreiber D, Wong AT, Schwartz D, Rineer J. Utilization of hyperfractionated radiation in small-cell lung cancer and its impact on survival. *J Thorac Oncol* (2015) 10(12):1770–5. doi:10.1097/JTO.0000000000000672
21. Overgaard J, Hansen HS, Overgaard M, Bastholt L, Berthelsen A, Specht L, et al. A randomized double-blind phase III study of nimorazole as a hypoxic radiosensitizer of primary radiotherapy in supraglottic larynx and pharynx carcinoma. Results of the Danish Head and Neck Cancer Study (DAHANCA) Protocol 5-85. *Radiother Oncol* (1998) 46(2):135–46. doi:10.1016/S0167-8140(97)00220-X
22. Kirkpatrick JP, Meyer JJ, Marks LB. The linear-quadratic model is inappropriate to model high dose per fraction effects in radiosurgery. *Semin Radiat Oncol* (2008) 18(4):240–3. doi:10.1016/j.semradonc.2008.04.005
23. Brenner DJ. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Semin Radiat Oncol* (2008) 18(4):234–9. doi:10.1016/j.semradonc.2008.04.004
24. Timmerman RD. An overview of hypofractionation and introduction to this issue of seminars in radiation oncology. *Semin Radiat Oncol* (2008) 18(4):215–22. doi:10.1016/j.semradonc.2008.04.001
25. Kirkpatrick JP, Soltys SG, Lo SS, Beal K, Shrieve DC, Brown PD. The radiosurgery fractionation quandary: single fraction or hypofractionation? *Neuro Oncol* (2017) 19(Suppl_2):ii38–49. doi:10.1093/neuonc/now301
26. Haviland JS, Owen JR, Dewar JA, Agrawal RK, Barrett J, Barrett-Lee PJ, et al. The UK Standardisation of Breast Radiotherapy (START) trials of radiotherapy hypofractionation for treatment of early breast cancer: 10-year follow-up results of two randomised controlled trials. *Lancet Oncol* (2013) 14(11):1086–94. doi:10.1016/S1470-2045(13)70386-3
27. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* (2001) 291(5507):1304–51. doi:10.1126/science.1058040
28. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* (2001) 409(6822):860–921. doi:10.1038/35057062
29. Tucker SL, Turesson I, Thames HD. Evidence for individual differences in the radiosensitivity of human skin. *Eur J Cancer* (1992) 28A(11):1783–91. doi:10.1016/0959-8049(92)90004-L
30. Bentzen SM, Overgaard M, Overgaard J. Clinical correlations between late normal tissue endpoints after radiotherapy: implications for predictive assays of radiosensitivity. *Eur J Cancer* (1993) 29A(10):1373–6. doi:10.1016/0959-8049(93)90004-Y
31. Safwat A, Bentzen SM, Turesson I, Hendry JH. Deterministic rather than stochastic factors explain most of the variation in the expression of skin telangiectasia after radiotherapy. *Int J Radiat Oncol Biol Phys* (2002) 52(1):198–204. doi:10.1016/S0360-3016(01)02690-6
32. Andreassen CN, Schack LM, Laursen LV, Alsner J. Radiogenomics – current status, challenges and future directions. *Cancer Lett* (2016) 382(1):127–36. doi:10.1016/j.canlet.2016.01.035
33. Andreassen CN. Searching for genetic determinants of normal tissue radiosensitivity – are we on the right track? *Radiother Oncol* (2010) 97(1):1–8. doi:10.1016/j.radonc.2010.07.018
34. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* (2002) 4(2):45–61. doi:10.1097/00125817-200203000-00002
35. Andreassen CN, Alsner J. Genetic variants and normal tissue toxicity after radiotherapy: a systematic review. *Radiother Oncol* (2009) 92(3):299–309. doi:10.1016/j.radonc.2009.06.015

36. West C, Rosenstein BS, Alsner J, Azria D, Barnett G, Begg A, et al. Establishment of a radiogenomics consortium. *Int J Radiat Oncol Biol Phys* (2010) 76(5):1295–6. doi:10.1016/j.ijrobp.2009.12.017
37. Rosenstein BS. Radiogenomics: identification of genomic predictors for radiation toxicity. *Semin Radiat Oncol* (2017) 27(4):300–9. doi:10.1016/j.semradonc.2017.04.005
38. Fachal L, Gomez-Caamano A, Barnett GC, Peleteiro P, Carballo AM, Calvo-Crespo P, et al. A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1. *Nat Genet* (2014) 46(8):891–4. doi:10.1038/ng.3020
39. Kerns SL, Dorling L, Fachal L, Bentzen S, Pharoah PD, Barnes DR, et al. Meta-analysis of genome wide association studies identifies genetic markers of late toxicity following radiotherapy for prostate cancer. *EBioMedicine* (2016) 10:150–63. doi:10.1016/j.ebiom.2016.07.022
40. Garber K. Oncologists await historic first: a pan-tumor predictive marker, for immunotherapy. *Nat Biotechnol* (2017) 35(4):297–8. doi:10.1038/nbt0417-297a
41. Coyne GO, Takebe N, Chen AP. Defining precision: the precision medicine initiative trials NCI-MPACT and NCI-MATCH. *Curr Probl Cancer* (2017) 41(3):182–93. doi:10.1016/j.currproblcancer.2017.02.001
42. Engelman JA, Janne PA. Mechanisms of acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in non-small cell lung cancer. *Clin Cancer Res* (2008) 14(10):2895–9. doi:10.1158/1078-0432.CCR-07-2248
43. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat Rev Cancer* (2012) 12(7):487–93. doi:10.1038/nrc3298
44. Mamounas EP, Tang G, Fisher B, Paik S, Shak S, Costantino JP, et al. Association between the 21-gene recurrence score assay and risk of locoregional recurrence in node-negative, estrogen receptor-positive breast cancer: results from NSABP B-14 and NSABP B-20. *J Clin Oncol* (2010) 28(10):1677–83. doi:10.1200/JCO.2009.23.7610
45. Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. *J Clin Oncol* (2008) 26(5):729–35. doi:10.1200/JCO.2007.14.3222
46. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* (2017) 23(6):703–13. doi:10.1038/nm.4333
47. Sottoriva A, Spiteri I, Piccirillo SG, Touloumis A, Collins VP, Marioni JC, et al. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A* (2013) 110(10):4009–14. doi:10.1073/pnas.1219747110
48. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. *Nat Genet* (2015) 47(3):209–16. doi:10.1038/ng.3214
49. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* (2010) 467(7319):1114–7. doi:10.1038/nature09515
50. Makohon-Moore A, Iacobuzio-Donahue CA. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nat Rev Cancer* (2016) 16(9):553–65. doi:10.1038/nrc.2016.66
51. Turajlic S, Swanton C. Metastasis as an evolutionary process. *Science* (2016) 352(6282):169–75. doi:10.1126/science.aaf2784
52. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* (2012) 366(10):883–92. doi:10.1056/NEJMoa1113205
53. El Naqa I, Kerns SL, Coates J, Luo Y, Speers C, West CML, et al. Radiogenomics and radiotherapy response modeling. *Phys Med Biol* (2017) 62(16):R179–206. doi:10.1088/1361-6560/aa7c55
54. Yard BD, Adams DJ, Chie EK, Tamayo P, Battaglia JS, Gopal P, et al. A genetic basis for the variation in the vulnerability of cancer to DNA damage. *Nat Commun* (2016) 7:11428. doi:10.1038/ncomms11428
55. Zhao SG, Chang SL, Spratt DE, Erho N, Yu M, Ashab HA, et al. Development and validation of a 24-gene predictor of response to postoperative radiotherapy in prostate cancer: a matched, retrospective analysis. *Lancet Oncol* (2016) 17(11):1612–20. doi:10.1016/S1470-2045(16)30491-0
56. Torres-Roca JF, Eschrich S, Zhao H, Bloom G, Sung J, McCarthy S, et al. Prediction of radiation sensitivity using a gene expression classifier. *Cancer Res* (2005) 65(16):7169–76. doi:10.1158/0008-5472.CAN-05-0656
57. Eschrich SA, Fulp WJ, Pawitan Y, Foekens JA, Smid M, Martens JW, et al. Validation of a radiosensitivity molecular signature in breast cancer. *Clin Cancer Res* (2012) 18(18):5134–43. doi:10.1158/1078-0432.CCR-12-0891
58. Scott JG, Berglund A, Schell MJ, Mihaylov I, Fulp WJ, Yue B, et al. A genome-based model for adjusting radiotherapy dose (GARD): a retrospective, cohort-based study. *Lancet Oncol* (2017) 18(2):202–11. doi:10.1016/S1470-2045(16)30648-9
59. Bishop CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. New York, NY: Springer-Verlag New York, Inc (2006).
60. Kang J, Schwartz R, Flickinger J, Beriwal S. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. *Int J Radiat Oncol Biol Phys* (2015) 93(5):1127–35. doi:10.1016/j.ijrobp.2015.07.2286
61. Coates J, Souhami L, El Naqa I. Big data analytics for prostate radiotherapy. *Front Oncol* (2016) 6:149. doi:10.3389/fonc.2016.00149
62. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.
63. Mathworks. *MATLAB: Statistics and Machine Learning Toolbox*. Natick, MA: MathWorks (2018).
64. Team RC. *R: A Language and Environment for Statistical Computing*. Auckland: R Core Team (2013).
65. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci* (2001) 16(3):199–231. doi:10.1214/ss/1009213725
66. Shmueli G. To explain or to predict? *Stat Sci* (2010) 25(3):289–310. doi:10.1214/10-STS330
67. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, et al. Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* (2006) 355(6):560–9. doi:10.1056/NEJMoa052933
68. Satija A, Yu E, Willett WC, Hu FB. Understanding nutritional epidemiology and its role in policy. *Adv Nutr* (2015) 6(1):5–18. doi:10.3945/an.114.007492
69. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* (2015) 68(9):1046–58. doi:10.1016/j.jclinepi.2015.05.029
70. Saeys Y, Abeel T, Van de Peer Y, editors. *Robust Feature Selection Using Ensemble Feature Selection Techniques. Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer (2008).
71. Nie F, Huang H, Cai X, Ding C. Efficient and robust feature selection via joint l2,1-norms minimization. *Proceedings of the 23rd International Conference on Neural Information Processing Systems*. (Vol. 2), Vancouver, BC: Curran Associates Inc (2010). p. 1813–21. 2997098.
72. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* (2015) 16(6):321–32. doi:10.1038/nrg3920
73. Ng AY, Jordan MI, editors. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: *NIPS'01 Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. Vancouver, BC (2002).
74. Valdes G, Luna JM, Eaton E, Simone CB II, Ungar LH, Solberg TD. MediBoost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Sci Rep* (2016) 6:37854. doi:10.1038/srep37854
75. Bellman R. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press (1961).
76. Noble WS. What is a support vector machine? *Nat Biotechnol* (2006) 24(12):1565–7. doi:10.1038/nbt1206-1565
77. Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nat Rev Genet* (2014) 15(11):722–33. doi:10.1038/nrg3747
78. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology – multifactorial decision support systems. *Nat Rev Clin Oncol* (2013) 10(1):27–40. doi:10.1038/nrclinonc.2012.196
79. El Naqa I, Li R, Murphy MJ, editors. *Machine Learning in Radiation Oncology: Theory and Applications*. 1 ed. New York, NY: Springer International Publishing (2015).

80. Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, et al. Estimating dataset size requirements for classifying DNA microarray data. *J Comput Biol* (2003) 10(2):119–42. doi:10.1089/106652703321825928
81. Valdes G, Solberg TD, Heskel M, Ungar L, Simone CB II. Using machine learning to predict radiation pneumonitis in patients with stage I non-small cell lung cancer treated with stereotactic body radiation therapy. *Phys Med Biol* (2016) 61(16):6105–20. doi:10.1088/0031-9155/61/16/6105
82. Schwartz R. *Biological Modeling and Simulation: A Survey of Practical Models, Algorithms, and Numerical Methods*. Cambridge, MA: MIT Press (2008). xii, 389 p.
83. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* (2002) 347(25):1999–2009. doi:10.1056/NEJMoa021967
84. Institute of Medicine. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington, DC: National Academies Press (2012). doi:10.17226/13297
85. Kolata G. *How Bright Promise in Cancer Testing Fell Apart*. New York, NY: The New York Times (2011).
86. Goldberg P. Duke officials silenced med student who reported trouble in Anil Potti's Lab. *Cancer Lett* (2015) 40(1):3.
87. Freedman DA. A note on screening regression equations. *Am Stat* (1983) 37(2):152–5. doi:10.1080/00031305.1983.10482729
88. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. *Chemometr Intell Lab Syst* (2006) 84(1):69–74. doi:10.1016/j.chemolab.2006.04.021
89. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* (2010) 11:2079–107.
90. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet* (2014) 10(11):e1004754. doi:10.1371/journal.pgen.1004754
91. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* (2003) 3:1157–82.
92. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY: Springer New York Inc (2001).
93. Bush WS, Moore JH. Chapter 11: genome-wide association studies. *PLoS Comput Biol* (2012) 8(12):e1002822. doi:10.1371/journal.pcbi.1002822
94. Yang P, Ho JW, Yang YH, Zhou BB. Gene-gene interaction filtering with ensemble of filters. *BMC Bioinformatics* (2011) 12(Suppl 1):S10. doi:10.1186/1471-2105-12-S1-S10
95. Moore JH. Epistasis analysis using ReliefF. *Methods Mol Biol* (2015) 1253:315–25. doi:10.1007/978-1-4939-2155-3_17
96. Greene CS, Penrod NM, Kiralis J, Moore JH. Spatially uniform reliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min* (2009) 2(1):5. doi:10.1186/1756-0381-2-5
97. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell* (1997) 97(1–2):273–324. doi:10.1016/S0004-3702(97)00043-X
98. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol* (1996) 58(1):267–88.
99. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* (1970) 12(1):55–67. doi:10.1080/00401706.1970.10488635
100. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* (2005) 67(2):301–20. doi:10.1111/j.1467-9868.20005.00503.x
101. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Series B Stat Methodol* (2010) 72(4):417–73. doi:10.1111/j.1467-9868.2010.00740.x
102. Duda RO, Hart PE, Stork DG. *Pattern Classification*. 2nd ed. New York, NY: Wiley-Interscience (2000).
103. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* (2006) 2(12):e190. doi:10.1371/journal.pgen.0020190
104. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* (2006) 38(8):904–9. doi:10.1038/ng1847
105. Lee H, Grosse R, Ranganath R, Ng AY, editors. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal: ACM (2009).
106. Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S, editors. Recurrent neural network based language model. *Eleventh Annual Conference of the International Speech Communication Association*. Makuhari: International Speech Communication Association (2010).
107. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* (2015) 521(7553):436–44. doi:10.1038/nature14539
108. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. (2014). arXiv preprint arXiv:14091556.
109. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE (2016).
110. Ferrucci D, editor. Build Watson: an overview of DeepQA for the Jeopardy! Challenge. *2010 19th International Conference on Parallel Architectures and Compilation Techniques (PACT)*. Vienna: IEEE (2010).
111. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* (2017) 542(7639):115–8. doi:10.1038/nature21056
112. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* (2016) 316(22):2402–10. doi:10.1001/jama.2016.17216
113. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. (2017). arXiv preprint arXiv:171105225.
114. Somashekhar SP, Sepulveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, et al. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol* (2018) 29(2):418–23. doi:10.1093/annonc/mdx781
115. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* (2016) 278(2):563–77. doi:10.1148/radiol.2015151169
116. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng* (2017) 19(1):221–48. doi:10.1146/annurev-bioeng-071516-044442
117. Carlos RC, Kahn CE, Halabi S. Data science: big data, machine learning, and artificial intelligence. *J Am Coll Radiol* (2018) 15(3):497–8.
118. Choi W, Oh JH, Riyahi S, Liu CJ, Jiang F, Chen W, et al. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. *Med Phys* (2018) 45(4):1537–49. doi:10.1002/mp.12820
119. Crispin-Ortuzar M, Apte A, Grkovski M, Oh JH, Lee NY, Schoder H, et al. Predicting hypoxia status using a combination of contrast-enhanced computed tomography and [(18)F]-fluorodeoxyglucose positron emission tomography radiomics features. *Radiother Oncol* (2018) 127(1):36–42. doi:10.1016/j.radonc.2017.11.025
120. Coroller TP, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee SW, et al. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol* (2016) 119(3):480–6. doi:10.1016/j.radonc.2016.04.004
121. Coroller TP, Bi WL, Huynh E, Abedalthagafi M, Aizer AA, Greenwald NF, et al. Radiographic prediction of meningioma grade by semantic and radiomic features. *PLoS One* (2017) 12(11):e0187908. doi:10.1371/journal.pone.0187908
122. Leung MK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics* (2014) 30(12):i121–9. doi:10.1093/bioinformatics/btu277
123. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* (2015) 347(6218):1254806. doi:10.1126/science.1254806
124. Poplin R, Newburger D, Dijamco J, Nguyen N, Loy D, Gross SS, et al. Creating a universal SNP and small indel variant caller with deep neural networks. *BioRxiv* (2017):092890.
125. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* (2015) 12(10):931–4. doi:10.1038/nmeth.3547
126. Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, et al. Machine learning in genome-wide association studies. *Genet Epidemiol* (2009) 33(Suppl 1):S51–7. doi:10.1002/gepi.20473
127. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ* (2001) 322(7280):226–31. doi:10.1136/bmj.322.7280.226

128. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* (2010) 11:724. doi:10.1186/1471-2164-11-724
129. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* (1995) 57(1):289–300.
130. Shaffer JP. Multiple hypothesis testing. *Annu Rev Psychol* (1995) 46(1):561–84. doi:10.1146/annurev.ps.46.020195.003021
131. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *Am Stat* (2016) 70(2):129–33. doi:10.1080/00031305.2016.1154108
132. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* (2017) 101(1):5–22. doi:10.1016/j.ajhg.2017.06.005
133. Maher B. Personal genomes: the case of the missing heritability. *Nature* (2008) 456(7218):18–21. doi:10.1038/456018a
134. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* (2013) 14(6):379–89. doi:10.1038/nrg3472
135. Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. Two-stage designs for gene-disease association studies. *Biometrics* (2002) 58(1):163–70. doi:10.1111/j.0006-341X.2002.00163.x
136. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol* (2007) 31(7):776–88. doi:10.1002/gepi.20240
137. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* (2006) 38(2):209–13. doi:10.1038/ng1706
138. Molinaro AM, Carriero N, Bjornson R, Hartge P, Rothman N, Chatterjee N. Power of data mining methods to detect genetic associations and interactions. *Hum Hered* (2011) 72(2):85–97. doi:10.1159/000330579
139. Cortes C, Jackel LD, Solla SA, Vapnik V, Denker JS, editors. Learning curves: asymptotic values and rate of convergence. In: *Advances in Neural Information Processing Systems*. Denver, CO (1994). p. 327–34.
140. Dietrich R, Opper M, Sompolinsky H. Statistical mechanics of support vector networks. *Phys Rev Lett* (1999) 82(14):2975. doi:10.1103/PhysRevLett.82.2975
141. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* (2009) 10(6):392–404. doi:10.1038/nrg2579
142. Fish AE, Capra JA, Bush WS. Are interactions between cis-regulatory variants evidence for biological epistasis or statistical artifacts? *Am J Hum Genet* (2016) 99(4):817–30. doi:10.1016/j.ajhg.2016.07.022
143. Hemani G, Theocharidis A, Wei W, Haley C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* (2011) 27(11):1462–5. doi:10.1093/bioinformatics/btr172
144. Yung LS, Yang C, Wan X, Yu W. GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics* (2011) 27(9):1309–10. doi:10.1093/bioinformatics/btr114
145. Lucas G, Lluis-Ganella C, Subirana I, Musameh MD, Gonzalez JR, Nelson CP, et al. Hypothesis-based analysis of gene-gene interactions and risk of myocardial infarction. *PLoS One* (2012) 7(8):e41730. doi:10.1371/journal.pone.0041730
146. Bell JT, Timpson NJ, Rayner NW, Zeggini E, Frayling TM, Hattersley AT, et al. Genome-wide association scan allowing for epistasis in type 2 diabetes. *Ann Hum Genet* (2011) 75(1):10–9. doi:10.1111/j.1469-1809.2010.00629.x
147. Li J, Horstman B, Chen Y. Detecting epistatic effects in association studies at a genomic level based on an ensemble approach. *Bioinformatics* (2011) 27(13):i222–9. doi:10.1093/bioinformatics/btr227
148. Yoshida M, Koike A. SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinformatics* (2011) 12:469. doi:10.1186/1471-2105-12-469
149. Culverhouse RC. A comparison of methods sensitive to interactions with small main effects. *Genet Epidemiol* (2012) 36(4):303–11. doi:10.1002/gepi.21622
150. De Lobel L, Geurts P, Baele G, Castro-Giner F, Kogevinas M, Van Steen K. A screening methodology based on random forests to improve the detection of gene-gene interactions. *Eur J Hum Genet* (2010) 18(10):1127–32. doi:10.1038/ejhg.2010.48
151. Lin HY, Chen YA, Tsai YY, Qu X, Tseng TS, Park JY. TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. *Ann Hum Genet* (2012) 76(1):53–62. doi:10.1111/j.1469-1809.2011.00692.x
152. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* (2009) 25(6):714–21. doi:10.1093/bioinformatics/btp041
153. Wasserman L, Roeder K. High dimensional variable selection. *Ann Stat* (2009) 37(5A):2178–201. doi:10.1214/08-AOS646
154. Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genet Epidemiol* (2010) 34(3):275–85. doi:10.1002/gepi.20459
155. Schwarz DF, Konig IR, Ziegler A. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* (2010) 26(14):1752–8. doi:10.1093/bioinformatics/btq257
156. Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, et al. SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics* (2012) 13:164. doi:10.1186/1471-2105-13-164
157. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* (2001) 69(1):138–47. doi:10.1086/321276
158. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet* (2007) 39(9):1167–73. doi:10.1038/ng2110
159. Meinshausen N, Meier L, Bühlmann P. p-Values for high-dimensional regression. *J Am Stat Assoc* (2009) 104(488):1671–81. doi:10.1198/jasa.2009.tm08647
160. Mieth B, Kloft M, Rodriguez JA, Sonnenburg S, Vobruba R, Morcillo-Suarez C, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci Rep* (2016) 6:36671. doi:10.1038/srep36671
161. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Test* (2003) 12(1):1–77. doi:10.1007/BF02595811
162. Nguyen TT, Huang J, Wu Q, Nguyen T, Li M. Genome-wide association data classification and SNPs selection using two-stage quality-based random forests. *BMC Genomics* (2015) 16(Suppl 2):S5. doi:10.1186/1471-2164-16-S2-S5
163. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Res* (2011) 39(9):e62. doi:10.1093/nar/gkr064
164. Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, Rao DC. Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genet Epidemiol* (2011) 35(2):111–8. doi:10.1002/gepi.20556
165. Oh JH, Kerns S, Ostrer H, Powell SN, Rosenstein B, Deasy JO. Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. *Sci Rep* (2017) 7:43381. doi:10.1038/srep43381
166. Lee S, Kerns S, Ostrer H, Rosenstein B, Deasy JO, Oh JH. Machine learning on a genome-wide association study to predict late genitourinary toxicity following prostate radiotherapy. *Int J Radiat Oncol Biol Phys* (2018) 101(1):128–35. doi:10.1016/j.ijrobp.2018.01.054
167. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press (2009).
168. Murphy K. *Learning Bayes Net Structure from Sparse Data Sets*. Technical report. Berkeley: Comp. Sci. Div., UC (2001).
169. Lee S, Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. *Med Phys* (2015) 42(5):2421–30. doi:10.1118/1.4915284
170. Luo Y, El Naqa I, McShan DL, Ray D, Lohse I, Matuszak MM, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother Oncol* (2017) 123(1):85–92. doi:10.1016/j.radonc.2017.02.004
171. Mourad R, Sinoquet C, Leray P. A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. *BMC Bioinformatics* (2011) 12:16. doi:10.1186/1471-2105-12-16
172. Breiman L. Random forests. *Mach Learn* (2001) 45(1):5–32. doi:10.1023/A:1010933404324
173. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of random forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet* (2010) 11:49. doi:10.1186/1471-2156-11-49
174. Cosgun E, Limdi NA, Duarte CW. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with

application to warfarin dose prediction in African Americans. *Bioinformatics* (2011) 27(10):1384–9. doi:10.1093/bioinformatics/btr159
175. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* (2003) 19(13):1636–43. doi:10.1093/bioinformatics/btg210
176. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* (2006) 7:3. doi:10.1186/1471-2105-7-3
177. Boulesteix AL, Porzelius C, Daumer M. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics* (2008) 24(15):1698–706. doi:10.1093/bioinformatics/btn262
178. Segal MR. *Machine Learning Benchmarks and Random Forest Regression*. Netherlands: Kluwer Academic Publishers (2004).
179. Liaw A, Wiener M. Classification and regression by random forest. *R News* (2002) 2(3):18–22.
180. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet* (2004) 5:32. doi:10.1186/1471-2156-5-32
181. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* (2011) 27(14):1986–94. doi:10.1093/bioinformatics/btr300
182. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics* (2010) 11:110. doi:10.1186/1471-2105-11-110
183. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* (2007) 8:25. doi:10.1186/1471-2105-8-25
184. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* (2006) 15(3):651–74. doi:10.1198/106186006X133933
185. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in the presence of population structure. *Nat Commun* (2015) 6:7432. doi:10.1038/ncomms8432
186. Paul D, Bair E, Hastie T, Tibshirani R. "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann Stat* (2008) 36(4):1595–618. doi:10.1214/009053607000000578
187. Platt J. Probabilities for SV Machines. In: Smola AJ, Bartlett PL, Schlköpf B, Schuurmans D, editors. *Advances in Large Margin Classifiers*. Cambridge, MA, London, England: The MIT Press (2000). p. 61–74.
188. Wang WA, Lai LC, Tsai MH, Lu TP, Chuang EY. Development of a prediction model for radiosensitivity using the expression values of genes and long non-coding RNAs. *Oncotarget* (2016) 7(18):26739–50. doi:10.18632/oncotarget.8496
189. Nimeus-Malmstrom E, Krogh M, Malmstrom P, Strand C, Fredriksson I, Karlsson P, et al. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence after breast-conservation surgery, with or without postoperative radiotherapy. *Breast Cancer Res* (2008) 10(2):R34. doi:10.1186/bcr1997
190. Hayashida Y, Honda K, Osaka Y, Hara T, Umaki T, Tsuchida A, et al. Possible prediction of chemoradiosensitivity of esophageal cancer by serum protein profiling. *Clin Cancer Res* (2005) 11(22):8042–7. doi:10.1158/1078-0432.CCR-05-0656
191. Gaspar P, Carbonell J, Oliveira JL. On the parameter optimization of support vector machines for binary classification. *J Integr Bioinform* (2012) 9(3):201. doi:10.2390/biecoll-jib-2012-201
192. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Exp Syst Appl* (2009) 36(2, Pt 2):3240–7. doi:10.1016/j.eswa.2008.01.009
193. Trainor PJ, DeFilippis AP, Rai SN. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites* (2017) 7(2):E30. doi:10.3390/metabo7020030
194. El Naqa I, Bradley JD, Lindsay PE, Hope AJ, Deasy JO. Predicting radiotherapy outcomes using statistical learning techniques. *Phys Med Biol* (2009) 54(18):S9–30. doi:10.1088/0031-9155/54/18/S02
195. Elwood M. Proteomic patterns in serum and identification of ovarian cancer. *Lancet* (2002) 360(9327):170; author reply–1. doi:10.1016/S0140-6736(02)09389-3
196. Pearl DC. Proteomic patterns in serum and identification of ovarian cancer. *Lancet* (2002) 360(9327):169–70; author reply 70–1. doi:10.1016/S0140-6736(02)09388-1
197. Rockhill B. Proteomic patterns in serum and identification of ovarian cancer. *Lancet* (2002) 360(9327):169; author reply 70–1. doi:10.1016/S0140-6736(02)09387-X
198. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* (2002) 16:321–57.
199. Cho BH, Yu H, Lee J, Chee YJ, Kim IY, Kim SI. Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels. *IEEE Trans Inf Technol Biomed* (2008) 12(2):247–56. doi:10.1109/TITB.2007.902300
200. Van Belle V, Van Calster B, Van Huffel S, Suykens JA, Lisboa P. Explaining support vector machines: a color based nomogram. *PLoS One* (2016) 11(10):e0164568. doi:10.1371/journal.pone.0164568
201. Cooper GF, Herskovits E. A Bayesian method for constructing Bayesian belief networks from databases. In: D'Ambrosio BD, Smets P, Bonissone PP, editors. *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann Publishers, Inc. (1991). p. 86–94.
202. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci U S A* (2003) 100(14):8348–53. doi:10.1073/pnas.0832373100
203. Oh JH, Craft J, Al Lozi R, Vaidya M, Meng Y, Deasy JO, et al. A Bayesian network approach for modeling local failure in lung cancer. *Phys Med Biol* (2011) 56(6):1635–51. doi:10.1088/0031-9155/56/6/008
204. Liu J, Page D, Nassif H, Shavlik J, Peissig P, McCarty C, et al. Genetic variants improve breast cancer risk prediction on mammograms. *AMIA Annu Symp Proc* (2013) 2013:876–85.
205. Lee S, Jiang X. Modeling miRNA-mRNA interactions that cause phenotypic abnormality in breast cancer patients. *PLoS One* (2017) 12(8):e0182666. doi:10.1371/journal.pone.0182666
206. Wang W, Baladandayuthapani V, Holmes CC, Do KA. Integrative network-based Bayesian analysis of diverse genomics data. *BMC Bioinformatics* (2013) 14(Suppl 13):S8. doi:10.1186/1471-2105-14-S13-S8
207. Prestat E, de Morais SR, Vendrell JA, Thollet A, Gautier C, Cohen PA, et al. Learning the local Bayesian network structure around the ZNF217 oncogene in breast tumours. *Comput Biol Med* (2013) 43(4):334–41. doi:10.1016/j.compbiomed.2012.12.002
208. Mattina J, Carlisle B, Hachem Y, Fergusson D, Kimmelman J. Inefficiencies and patient burdens in the development of the targeted cancer drug sorafenib: a systematic review. *PLoS Biol* (2017) 15(2):e2000487. doi:10.1371/journal.pbio.2000487
209. Roviello G, Bachelot T, Hudis CA, Curigliano G, Reynolds AR, Petrioli R, et al. The role of bevacizumab in solid tumours: a literature based meta-analysis of randomised trials. *Eur J Cancer* (2017) 75:245–58. doi:10.1016/j.ejca.2017.01.026
210. Kimmelman J, Carlisle B, Gonen M. Drug development at the portfolio level is important for policy, care decisions and human protections. *JAMA* (2017) 318(11):1003–4. doi:10.1001/jama.2017.11502
211. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. STATISTICS. The reusable holdout: preserving validity in adaptive data analysis. *Science* (2015) 349(6248):636–8. doi:10.1126/science.aaa9375
212. Dwork C, editor. *Differential Privacy: A Survey of Results. International Conference on Theory and Applications of Models of Computation*. Xi'an: Springer (2008).
213. Narayanan A, Shmatikov V, editors. Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (SP 2008)*. Oakland: IEEE (2008).
214. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth AL, editors. Preserving statistical validity in adaptive data analysis. *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. Portland, OR: ACM (2015).
215. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A, editors. *Generalization in Adaptive Data Analysis and Holdout Reuse. Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press (2015).
216. Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat Genet* (2017) 49(9):1319–25. doi:10.1038/ng.3931

217. O'Callaghan ME, Raymond E, Campbell JM, Vincent AD, Beckmann K, Roder D, et al. Patient-reported outcomes after radiation therapy in men with prostate cancer: a systematic review of prognostic tool accuracy and validity. *Int J Radiat Oncol Biol Phys* (2017) 98(2):318–37. doi:10.1016/j.ijrobp.2017.02.024
218. Marks LB, Yorke ED, Jackson A, Ten Haken RK, Constine LS, Eisbruch A, et al. Use of normal tissue complication probability models in the clinic. *Int J Radiat Oncol Biol Phys* (2010) 76(3 Suppl):S10–9. doi:10.1016/j.ijrobp.2009.07.1754
219. Rosenstein BS, Capala J, Efstathiou JA, Hammerbacher J, Kerns SL, Kong FS, et al. How will big data improve clinical and basic research in radiation therapy? *Int J Radiat Oncol Biol Phys* (2016) 95(3):895–904. doi:10.1016/j.ijrobp.2015.11.009
220. Valentini V, Bourhis J, Hollywood D. ESTRO 2012 strategy meeting: vision for radiation oncology. *Radiother Oncol* (2012) 103(1):99–102. doi:10.1016/j.radonc.2012.03.010
221. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* (2002) 359(9306):572–7. doi:10.1016/S0140-6736(02)07746-2
222. Pollack A. *New Cancer Test Stirs Hope and Concern*. New York, NY: New York Times (2004). Sect. Science.
223. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* (2003) 4:24. doi:10.1186/1471-2105-4-24
224. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* (2004) 20(5):777–85. doi:10.1093/bioinformatics/btg484
225. Mor G, Visintin I, Lai Y, Zhao H, Schwartz P, Rutherford T, et al. Serum protein markers for early detection of ovarian cancer. *Proc Natl Acad Sci U S A* (2005) 102(21):7677–82. doi:10.1073/pnas.0502178102
226. Visintin I, Feng Z, Longton G, Ward DC, Alvero AB, Lai Y, et al. Diagnostic markers for early detection of ovarian cancer. *Clin Cancer Res* (2008) 14(4):1065–72. doi:10.1158/1078-0432.CCR-07-1569
227. Buchen L. Cancer: missing the mark. *Nature* (2011) 471(7339):428–32. doi:10.1038/471428a
228. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, et al. Genomic signatures to guide the use of chemotherapeutics. *Nat Med* (2006) 12(11):1294–300. doi:10.1038/nm1491
229. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, et al. Retraction: genomic signatures to guide the use of chemotherapeutics. *Nat Med* (2011) 17(1):135. doi:10.1038/nm0111-135
230. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann Appl Stat* (2009) 3(4):1309–34. doi:10.1214/09-AOAS291
231. Gatter K. FDA oversight of laboratory-developed tests: where are we now? *Arch Pathol Lab Med* (2017) 141(6):746–8. doi:10.5858/arpa.2017-0053-ED
232. Wallner PE, Anscher MS, Barker CA, Bassetti M, Bristow RG, Cha YI, et al. Current status and recommendations for the future of research, teaching, and testing in the biological sciences of radiation oncology: report of the American Society for Radiation Oncology Cancer Biology/Radiation Biology Task Force, executive summary. *Int J Radiat Oncol Biol Phys* (2014) 88(1):11–7. doi:10.1016/j.ijrobp.2013.09.040
233. Steinberg M, McBride WH, Vlashi E, Pajonk F. National Institutes of Health funding in radiation oncology: a snapshot. *Int J Radiat Oncol Biol Phys* (2013) 86(2):234–40. doi:10.1016/j.ijrobp.2013.01.030
234. Wallner PE, Ang KK, Zietman AL, Harris JR, Ibbott GS, Mahoney MC, et al. The American Board of Radiology Holman Research Pathway: 10-year retrospective review of the program and participant performance. *Int J Radiat Oncol Biol Phys* (2013) 85(1):29–34. doi:10.1016/j.ijrobp.2012.04.024
235. Formenti SC, Bonner JF, Hahn SM, Lawrence TS, Liu FF, Thomas CR Jr. Raising the next generation of physician-scientists: the chairs' perspective. *Int J Radiat Oncol Biol Phys* (2015) 92(2):211–3. doi:10.1016/j.ijrobp.2015.01.038

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

REVIEW
published: 27 July 2018
doi: 10.3389/fonc.2018.00266

The Role of Machine Learning in Knowledge-Based Response-Adapted Radiotherapy

Huan-Hsin Tseng, Yi Luo, Randall K. Ten Haken and Issam El Naqa*

Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, United States

With the continuous increase in radiotherapy patient-specific data from multimodality imaging and biotechnology molecular sources, knowledge-based response-adapted radiotherapy (KBR-ART) is emerging as a vital area for radiation oncology personalized treatment. In KBR-ART, planned dose distributions can be modified based on observed cues in patients' clinical, geometric, and physiological parameters. In this paper, we present current developments in the field of adaptive radiotherapy (ART), the progression toward KBR-ART, and examine several applications of static and dynamic machine learning approaches for realizing the KBR-ART framework potentials in maximizing tumor control and minimizing side effects with respect to individual radiotherapy patients. Specifically, three questions required for the realization of KBR-ART are addressed: (1) what knowledge is needed; (2) how to estimate RT outcomes accurately; and (3) how to adapt optimally. Different machine learning algorithms for KBR-ART application shall be discussed and contrasted. Representative examples of different KBR-ART stages are also visited.

Keywords: adaptive radiotherapy, personalized treatment, deep learning, statistical learning, big data

OPEN ACCESS

Edited by:
John Varlotto, University of Massachusetts Medical School, United States

Reviewed by:
Hengyong Yu, University of Massachusetts Lowell, United States
Yidong Yang, University of Miami, United States

****Correspondence:***
Huan-Hsin Tseng
thuanhsi@med.umich.edu

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *25 February 2018*
Accepted: *27 June 2018*
Published: *27 July 2018*

Citation:
Tseng H-H, Luo Y, Ten Haken RK and El Naqa I (2018) The Role of Machine Learning in Knowledge-Based Response-Adapted Radiotherapy. Front. Oncol. 8:266. doi: 10.3389/fonc.2018.00266

1. INTRODUCTION

Recent advances in cancer multimodality imaging (CT/PET/MRI/US) and biotechnology (genomics, transcriptomics, proteomics, etc.) have resulted in tremendous growth in patient-specific information in radiation oncology, ushering in the new era of Big Data in radiotherapy. With the availability of the individual-specific data, such as clinical, dosimetric, imaging, molecular markers, before and/or during radiotherapy (RT) courses, new opportunities are becoming available for personalized radiotherapy treatment (1, 2).

The synthesis of this information into actionable knowledge to improve patient outcomes is currently a major goal of modern radiotherapy (RT). Subsequently, knowledge-based response-adapted radiotherapy (KBR-ART) has emerged as an important framework that aims to develop personalized treatments by adjusting dose distributions according to clinical, geometrical changes, and physiological parameters observed during a radiotherapy treatment course. The notion of KBR-ART extends the traditional concept of adapted RT (ART) (3, 4), primarily based on imaging information for guidance, into a more general ART framework that can receive and process all relevant patient-specific signals that can be useful for adaptive decision-making. Our goal is to explore in more details the processes involved in the KBR-ART framework that would allow aggregating and analyzing relevant patient information in a systematic manner to achieve more accurate decision making and optimize long-term outcomes.

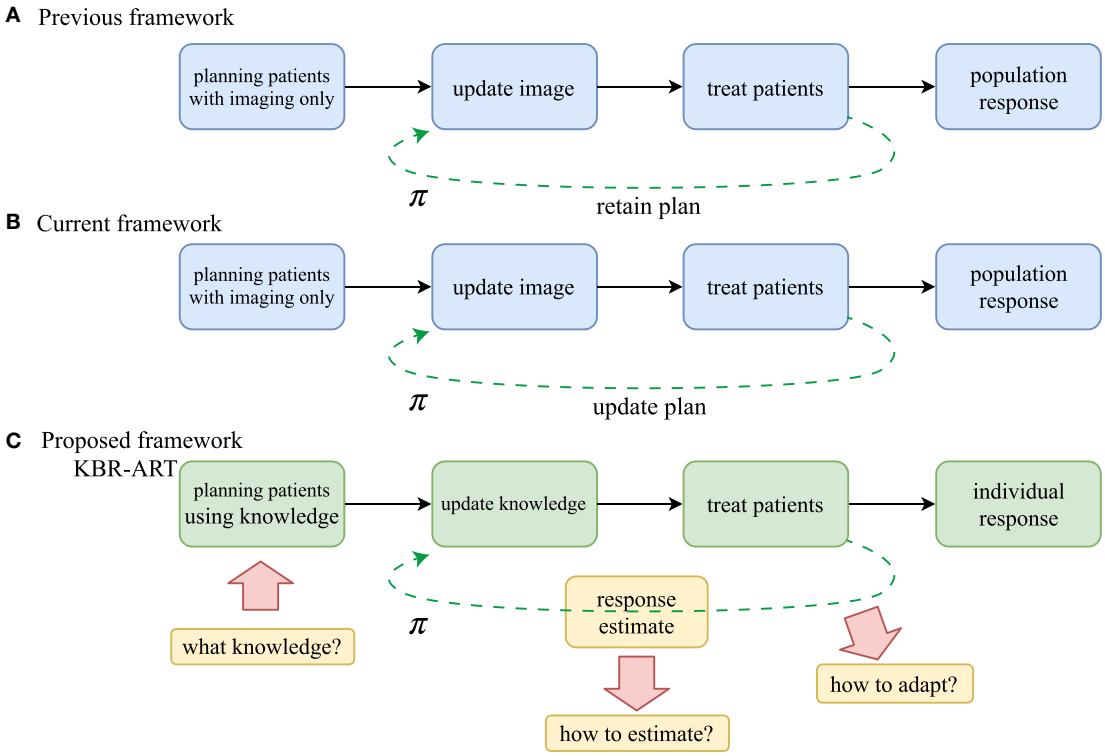

FIGURE 1 | Comparison of workflow of **(A)** non-adaptive RT, **(B)** current image-based ART, and **(C)** the proposed KBR-ART approach. The current ART **(B)** mostly relies on image guidance such as computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). In KBR-ART, the planning patients stage can utilize general knowledge about patient status (imaging + biological markers) as information for adapting treatment instead of using imaging only. Two major differences between previous/current RT and KBR-ART are that (1) knowledge is no longer restricted to imaging only and can include biological markers such as tumor genetics or blood-based inflammatory proteins (cytokines) to inform predictive modeling and decision-making; and (2) application process of machine learning for adapting a treatment plan π in KBR-ART.

The proposed KBR-ART framework can be thought of as being comprised of four stages, as depicted in **Figure 1**. These stages include: (1) *planning patients using available knowledge*, or pre-treatment modeling, (2) *updating the prediction models with evolving knowledge through the course of therapy*, or during-treatment modeling, (3) personalizing initial patient's treatments, and (4) *adapting the initial treatment to individual's responses*, where the two middle steps can be repeated at each radiation dose fraction (or few fractions) so that optimal treatment objectives are met and potentially long-term goals are optimized, i.e., long-term tumor control with limited side effects to surrounding normal tissues.

The first step in the implementation of a KBR-ART framework starts at the planning stage of patients by extending the current "image-only patients" into a more general preparation stage that can incorporate all relevant informatics signals for evaluating available treatment options, c.f. **Figures 1A,B**. Thus, the "K" in our KBR-ART refers to any useful knowledge (e.g., imaging (CT/PET/MRI) and biological markers (genomics, transcriptomics, proteomics, etc.) that can potentially aid the process of personalizing treatment to an individual patient's molecular characteristics and is not limited to imaging only as currently is the case. In Section 2, we shall introduce four major categories of data that are relevant to improved knowledge synthesis in RT. As the era of Big data (BD) is upon us, many useful tools applied for BD analytics are being actively developed in the context of modern machine learning algorithms, where KBR-ART is expected to be a prime beneficiary of this progress toward the development of dynamically personalized radiotherapy treatment leading to better outcomes and improved patients' quality of life. However, there are three essential questions pertaining to the successful development of a KBR-ART framework in radiotherapy that need to be addressed:

Q1: What knowledge should be synthesized for radiotherapy planning?
Q2: How can we develop powerful predictive outcome modeling techniques based on such knowledge?
Q3: How can we use these models in a strategically optimal manner to adapt a patient's treatment plan?

The answers to these three questions are at the core of successful development of the proposed KBR-ART framework and we shall attempt to address them in more depth in Sections 2–4 of this paper. During the process of exploring the answers to these questions, we shed more light on the pivotal role that machine learning algorithms play in the design and development of a modern KBR-ART system in subsequent sections as outlined in **Figure 2**.

A major inherent merit of the KBR-ART framework is that the treatment planning would be designed to dynamically adapt to

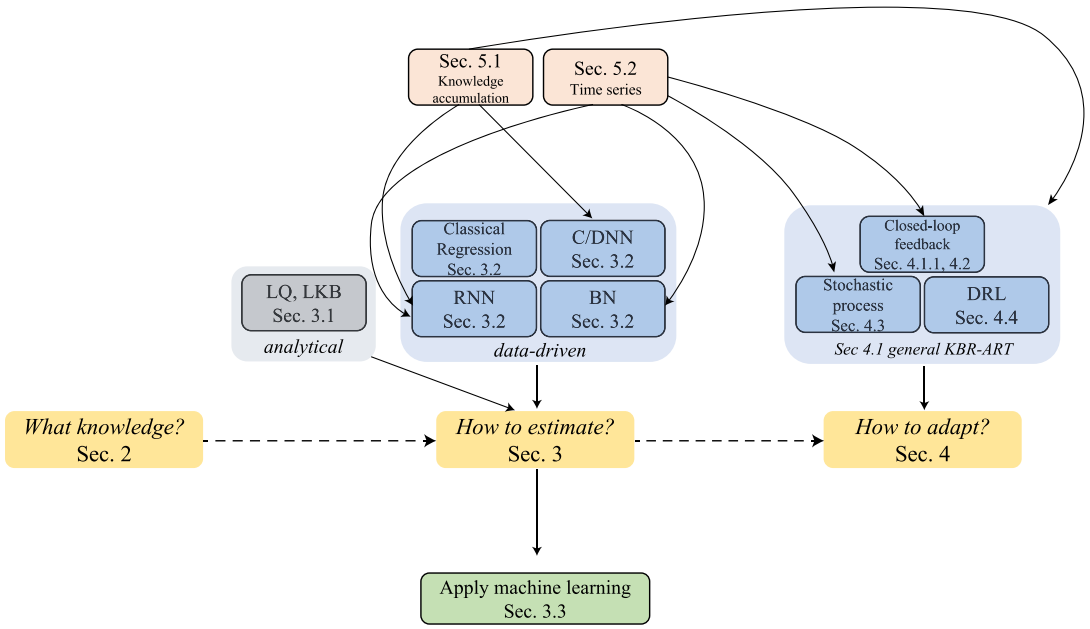

FIGURE 2 | The graph outlines the sectional relationships and the organization of this paper, where the main thread is through Sections 2–4 can address the three pertained questions involved in KBR-ART design and development, **Figure 1C**.

ongoing changes during the course of therapy to optimize radiotherapy goals of eradicating the tumor while minimizing harm to uninvolved normal tissue based on the individual patient's characteristics. As shown in **Figure 1**, adaptation of a treatment plan can be more formally accomplished in accordance to a decision making function π. This is represented in **Figure 1A** for the previous/current framework, where π is a non-varying function but in the case of KBR-ART, **Figure 1B**, π is a time-varying function that depends on the information (knowledge updates) available during the course of therapy. The following scenario may be used as an example on how KBR-ART can be implemented in practice: a given planned radiation course was considered optimal according to an initial population-based model such as traditional dose-based tumor control probability (TCP) and normal tissue complication probability (NTCP) and the goal is to optimize the uncomplicated tumor control $[p^+ = \mathrm{TCP} \cdot (1\text{–}\mathrm{NTCP})]$, for instance. Then, through the course of fractionated radiotherapy treatment, the patient did not achieve the predicted TCP value as expected, or worse suffered from unexpected toxicities due to treatment, i.e., NTCP exceeded the designed risk limit. This is where KBR-ART comes into action; to learn from current observations with its previous decisions taking into account available information during therapy and to adjust the course of action [e.g., increase dose to improve TCP or decrease it to specific organ-at-risk (OAR) to limit its NTCP] and develop a better personalized treatment plan based on the updated knowledge (from imaging and biomarkers) of the specific patient under treatment as shown in **Figure 1B**.

Much effort of this study will be devoted to tackling questions (ii and iii), which requires consideration of some advanced data-driven models that can also incorporate temporal information (i.e., knowledge updates). The steps involved in the development of a knowledge-adapted plan using the KBR-ART framework will be the main subject of this paper. For this purpose, we will first review pertained modern machine learning algorithms that feature modeling of sequential data. These include efficient *deep-learning* approaches such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and the more recently developed deep reinforcement learning (DRL). The subject of sequential data modeling have been applied in many diverse fields, such as handwriting recognition (5), speech recognition (6), bioinformatics (7), medical care (8, 9), and also high energy physics (10).

The introduced algorithms based on deep learning would require some basic background of neural networks (NNs) which are briefly reviewed in Section 3.2.2. Most of the notations in this paper are self-contained and self-consistent. In addition to the presented advanced data-driven models, we also provide probabilistic and statistical perspectives as a theoretical foundation for sequential machine learning models. In particular, via "*filtration*" we are to describe notions related to "knowledge accumulation" or "growing of knowledge" in more concrete manner. A main part of KBR-ART development relies on constructing a new RT plan prescription based on historical information; thus we would like to address issues related to representing knowledge accumulation in sequential learning models.

Moreover, we recognize that KBR-ART has a close analogy to stock pricing or autonomous car driving, in that it shares the same goal of analyzing acquired information a long a period of time to maximize final rewards (e.g., better radiotherapy treatment outcome in our case). Therefore, techniques derived from time series analysis will be helpful to analyze such sequential data from an analytical perspective, such as the trends and the stationarity of such stochastic (random) processes. In particular, it suffices for

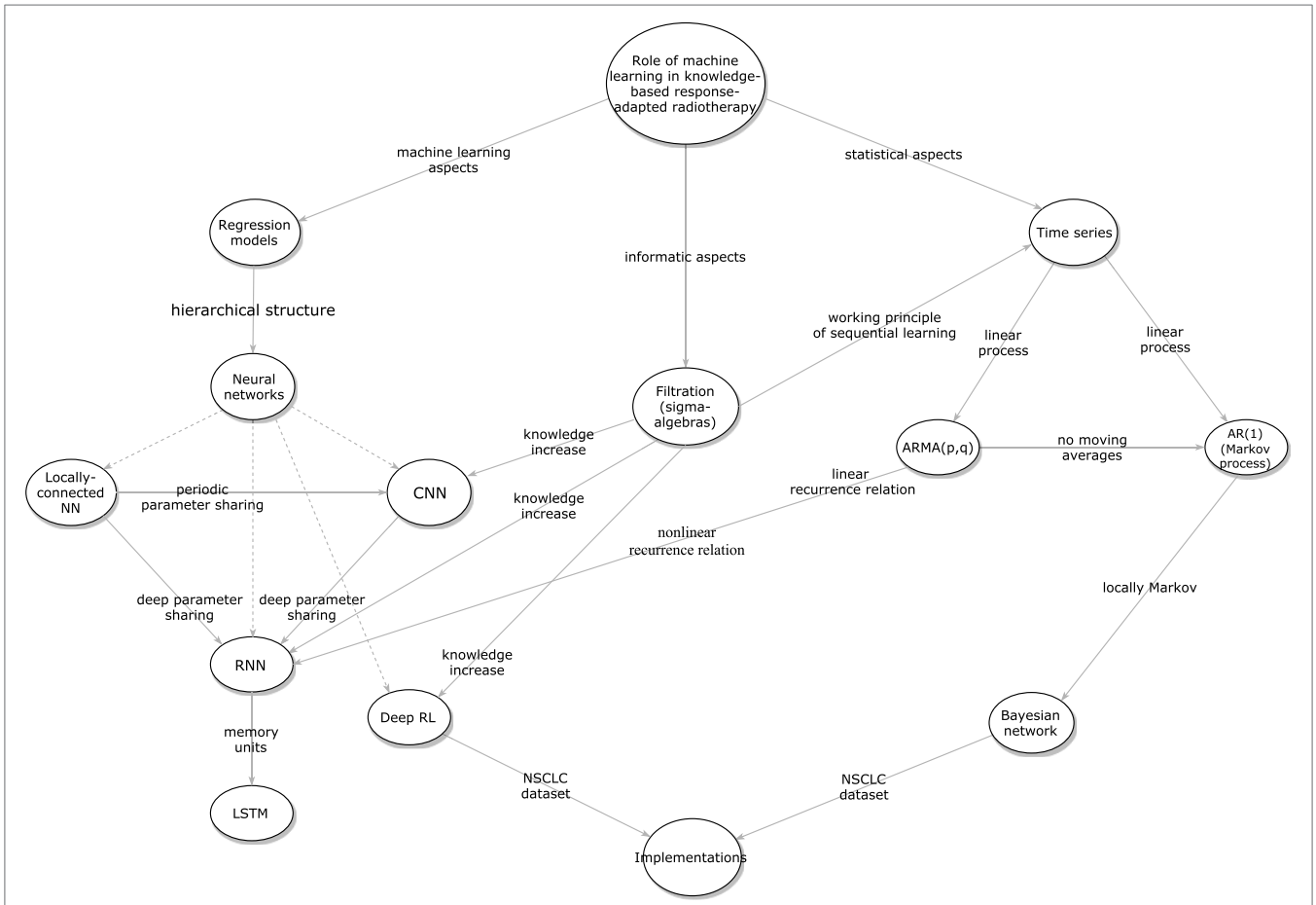

FIGURE 3 | The inter-relations between the different presented algorithms for designing KBR-ART a framework.

our purpose to revisit main *linear processes*, such as the autoregressive moving average (ARMA) model and its natural descendant the autoregressive (AR) models, which can be linked to Bayesian networks (BNs), another useful approach for dynamical learning as summarized in **Figure 3**. Together, our goal is to provide a comprehensive overview and a frontier survey that covers the major facets for the application of KBR-ART and layout the foundation for this emerging field.

It worth noticing that we organized the sections of this paper so that it follows the necessary building steps for the development of a successful KBR-ART framework as pertained to addressing the three aforementioned questions involved in KBR-ART implementation and review the related literature accordingly. Two implementations using non-small lung cancer (NSCLC) datasets will be presented for illustration.

2. Q1: WHAT KNOWLEDGE TO BE USED FOR KBR-ART PLANNING?

There are four major types of RT data that are potentially useful as part of the knowledge synthesis for KBR-ART: *clinical, dosimetric, imaging radiomics, and biological data*. To understand why and how they can be informative for assessing treatment outcomes, we provide a brief description about these four categories of data.

2.1. Clinical Data

Clinical data refers to cancer diagnostic characteristics (e.g., grade, stage, histology, site, etc.), physiological metrics (e.g., blood cell counts, heart/pulse rates, pulmonary measurements, etc.), and patient-related information (e.g., comorbidities, gender, age, etc.). Due to their nature, clinical data can usually be found in unstructured format such that can be challenging for extracting information directly. Therefore, machine learning techniques for natural language processing could be useful for transforming such data into structured format (e.g., tabulated) before further processing (11).

2.2. Dosimetric Data

Dosimetric data are informatic to the treatment planning process in RT, which includes simulated calculation of radiation dose using computed tomography (CT) imaging. In particular, dose–volume metrics obtained out of dose–volume histograms (DVHs) are extensively investigated for outcome modeling (12–16). Useful metrics are typically the volume receiving greater than or equal to a certain dose (Vx), the minimum dose to the hottest x% of the volume (Dx), mean, maximum, minimum dose, etc. (17). Notably, a dedicated software based on MATLAB™ called "DREES" can derive theses metrics automatically and apply them in outcome prediction models of RT response (18).

2.3. Radiomics Data

Radiomics is a field of medical imaging study that aims to extract meaningful quantitative features from medical images and relate this information to clinical and biological endpoints. The most common imaging modality is CT, which has been considered the standard for treatment planning in RT. Other imaging modalities used for improving treatment monitoring and prognosis in various cancer types are also used, such as positron emission tomography (PET), and magnetic imaging resonance (MRI). These modalities can be used individually or combined (19, 20).

2.4. Biological Data

According to (21) a biomarker is defined as *"a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention."* Measurements of biomarkers are typically based on tissue or fluid specimens, which are analyzed using molecular biology laboratory techniques (22) and have the following two categories according to their biochemical sources:

(a) *Exogenous biomarkers*: by injecting foreign substance into patients such as that used in molecular imaging and are used in radiomics applications.
(b) *Endogenous biomarkers*: there exists two subclasses within this category:
   (i) *Expression biomarkers*: changes measured in protein levels or gene expression.
   (ii) *Genetic biomarkers*: measuring variations between the underlying DNA genetic code and tumors or normal tissues.

2.5. Example: Aggregating Relevant Knowledge From a Lung Cancer Dataset

In this paper, we shall apply an institutional non-small cell lung cancer (NSCLC) dataset (23) as an example for implementation of KBR-ART. The first step is to collect relevant knowledge from such dataset that is suitable for the purposes of adapting radiotherapy treatment planning during a fractionated course. These data will be used subsequently for outcome modeling (TCP/NTCP) and plan adaptation as discussed later.

2.5.1. Data Description

The NSCLC dataset was recorded from NSCLC patients, where they have been treated on prospective protocols with standard and dose escalated fractionation under IRB approval (24). Collectively, 125 patients with relatively complete characteristics were selected for predicting TCP (local control) and NTCP (radiation pneumonitis of grade 2 or above (RP2)).

The dataset had over 250 features containing positron emission tomography (PET) imaging radiomics features, circulating inflammatory cytokines, single-nucleotide polymorphisms (SNPs), circulating microRNAs, clinical factors, and dosimetric variables before and during radiotherapy. All features were recorded at three time periods (at baseline, at 2 weeks of treatment, and at 4 weeks). However, certain features were collected only at baseline such as microRNAs and SNPs. Thus, the data for the purpose of KBR-ART can be represented as forming 3 time blocks:

$$N \text{ samples} \left\{ \left[\begin{array}{cccc|cccc|cccc} x_{11}^{(0)} & x_{12}^{(0)} & \dots & x_{1n}^{(0)} & x_{11}^{(1)} & x_{12}^{(1)} & \dots & x_{1n}^{(1)} & x_{11}^{(2)} & x_{12}^{(2)} & \dots & x_{1n}^{(2)} \\ x_{21}^{(0)} & x_{22}^{(0)} & \dots & x_{2n}^{(0)} & x_{21}^{(1)} & x_{22}^{(1)} & \dots & x_{2n}^{(1)} & x_{21}^{(2)} & x_{22}^{(2)} & \dots & x_{2n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N1}^{(0)} & x_{N2}^{(0)} & \dots & x_{Nn}^{(0)} & x_{N1}^{(1)} & x_{N2}^{(1)} & \dots & x_{Nn}^{(1)} & x_{N1}^{(2)} & x_{N2}^{(2)} & \dots & x_{Nn}^{(2)} \end{array}\right] \right\}, \quad (1)$$

where $x_{ij}^{(k)}$ denotes the value of the jth feature of patient i at time period k.

Values of mean tumor and lung doses were computed in their 2 Gy equivalents (EQD2) by using the linear-quadratic (LQ) model (Section 3.1.1) with $\alpha/\beta = 10\,Gy$, $4\,Gy$ for the tumor and the lung, respectively. Generalized equivalent uniform doses (gEUDs) with various a parameters were also calculated for gross tumor volumes (GTVs) and uninvolved lungs (lung volumes exclusive of GTVs).

3. Q2: HOW TO ESTIMATE RADIOTHERAPY OUTCOME MODELS FROM AGGREGATED KNOWLEDGE?

Radiotherapy outcome models are typically expressed in terms of tumor control probability (TCP) and normal tissue complication probability (NTCP) (25, 26). In principle, both TCP and NTCP may be evaluated using analytical and/or data-driven models. Though the former provides structural formulation, it can be incomplete and less accurate due to the complexity of radiobiological processes. On the other hand, data-driven models tend to learn empirically from the data observed, and thus they are capable of considering higher complexities and interactions of irradiation with the biological system. The trade-offs between analytical models and data-driven models can vary in terms of radiobiological understanding and prediction accuracy. In the following, we list examples, more detailed description on treatment outcome models can be found in (27).

3.1. Analytical Models

These models are generally based on simplified understanding of radiobiological processes and can provide a mechanistic formalism of radiation interactions with live tissue.

3.1.1. TCP

The most prevalent TCP models are based on the linear quadratic (LQ) model (28) parametrized by the radiosensitivity ratio α/β derived from clonogenic cell survival curves. The LQ model expresses the survival fraction (SF) after irradiation as follows:

$$\mathrm{SF} = e^{-\alpha D - \beta D^2}, \quad (2)$$

where $D \geq 0$ is the total delivered dose. For n fractions of dose d in uniformly delivered fractions is represented by:

$$\mathrm{SF} = e^{-n\left(\alpha d + \beta d^2\right)}. \quad (3)$$

Many types of TCP models were proposed (28) in the literature such as the birth-death (29) and the Poisson-based (30) models, which are expressed as:

$$\mathrm{TCP} = e^{-N\cdot e^{-n\left(\alpha d+\beta d^2\right)-t\ln 2/T_{\mathrm{pot}}}}, \tag{4}$$

where N is the initial number of colonogenic cells, and T_{pot} denotes the potential cell doubling time, with t as the time difference within the total treatment elapse T, the lag period before accelerated clonogenic repopulation begins.

3.1.2. NTCP

The most frequently used analytical model is the Lyman–Kutcher–Burman (LKB) model, which is a phenomenological approach (31). In the uniform dose case, NTCP is expressed by a gaussian integral (probit function):

$$\mathrm{NTCP}_{m,D_{50}}(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\,du, \qquad \left(x = \frac{D-D_{50}}{m\,D_{50}}\right), \tag{5}$$

where D_{50} is defined as the dose that corresponds to NTCP probability (curve in **Figure 4**) of 50% and m is a parameter tuning the shape of the NTCP curve. Typical trade-off between TCP and NTCP to achieve a therapeutic ratio is shown in **Figure 4**.

To account for dose inhomogeneities in developing TCP/NTCP models, the Equivalent Uniform Dose (EUD) (32) or Generalized EUD (gEUD) (33) are used. Mimicking a weighted sum of doses, gEUD is given by:

$$\mathrm{gEUD} = \sqrt[a]{\sum_{i=1}^{n} v_i\,D_i^a}, \tag{6}$$

where v_i is the fractional organ volume receiving dose D_i and a is a volume parameter that depends on the tissue type. An $a<0$ value will correspond to minimum dose effect, which is typically associated with tumor response. An $a>0$ value will correspond to maximum dose effect, which is typically associated with serial normal tissue architecture response, while an $a=1$ will correspond to mean dose effect, which is associated with parallel normal tissue architecture response.

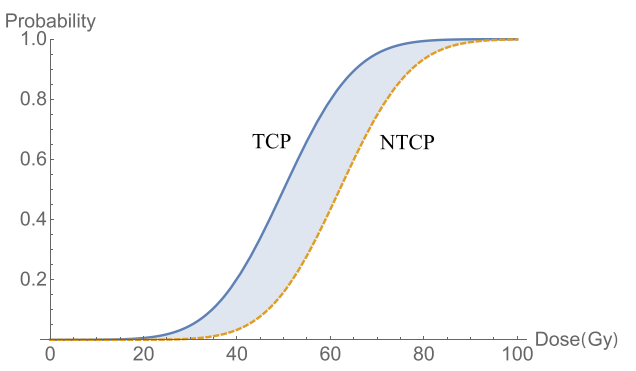

FIGURE 4 | An illustration of a therapeutic ratio showing that the trade-off between TCP and NTCP as delivered dose increases. The blue-shaded area between two curves TCP (blue) and NTCP (orange-dashed) is a best window for dose delivery.

More complex analytical models for toxicity can be developed by incorporating variables other than dose in the LKB model, for instance (34, 35):

$$\mathrm{NTCP}_{m,D_{50},\mathrm{DMFs}}(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\,du \tag{7}$$

with

$$x = \frac{D_{\mathrm{eff}}\cdot \mathrm{DMF}_1\cdot \mathrm{DMF}_2\cdots \mathrm{DMF}_k - D_{50}}{m\,D_{50}},$$

where the DMFs are dose modifying factors and represent the impact of covariates other than dose (e.g., single-nucleotide polymorphism (SNPs) genotype, copy number variations (CNVs), smoking status, etc.). Although analytical models are useful, in many circumstances, they are simply approximations of the complex physical and biological processes that are currently beyond such simple formalisms. Therefore, more data-driven approaches are being sought to achieve more accurate predictions of TCP/NTCP.

3.2. Data-Driven Models

By definition, data-driven models are approximations built based on observation of data. However, one drawback is that such modeling is likely not unique even from the same dataset and, therefore, one needs to choose a suitable technique that fits one's dataset best, which is an open question in the data science world. The purpose main of this section is to present several advanced data-driven techniques that can suite the implementation of predictive outcome modeling component of the KBR-ART framework. Below, we summarize some frequently used data-driven techniques for outcome modeling ranging from classical regression models to more advanced machine learning techniques.

3.2.1. Classical Models

Regression models such as Ridge, LASSO, and Logistic are commonly used foe building outcome models and follow conventional statistical approaches (36). They are essentially constructed by minimizing the following objective:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N}\left[\mathbf{y}_i - (\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\right]^2 + \lambda\cdot h(\mathbf{w}), \tag{8}$$

where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in \mathbb{R}$, $i=1,\ldots,N$, are the data input and outputs, respectively. Here, the weights $\mathbf{w} \in \mathbb{R}^n$ and bias $b \in \mathbb{R}$ are unknown parameters to be fitted by minimizing **regression error**, Equation (8). The second term in Equation (8) represents **penalty**, usually used to suppress possible model's overfitting. There are several types of penalty corresponding to different model characteristics, such as $h(\mathbf{w}) = \|\mathbf{w}\|$ is called the **LASSO** by Tibshirani (37), $h(\mathbf{w}) = \|\mathbf{w}\|^2$ is called the **Rigid (Tikhonov) regularization** (37), and $h(\mathbf{w}) = \lambda_1\|\mathbf{w}\| + \lambda_2\|\mathbf{w}\|^2$ is called the

Elastic Net regularization (38). The regularization parameter λ controls the magnitude of the penalty.

Due to the characteristic of L_1-norm, $||\cdot||$, the LASSO regularization tends to suppress many parameters to equal zero, so that the parameter vector is sparse, which makes it a natural candidate for relevant *feature selection* (39).

Another benefit of regression models other than their simplicity is the convex optimization property of their loss function, which guarantees optimal fitting parameters $\mathbf{w} = \mathbf{w}_*$. In fact, it can be explicitly solved using simple matrix inversion $\mathbf{w}_* = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \cdot \mathbf{X}^T\ \mathbf{y}$, for Ridge regression, for instance, where $\mathbf{X}$ is known from the given data:

$$\mathbf{X} \overset{def}{=} \begin{pmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ & \vdots & \\ - & \mathbf{x}_N & - \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nn} \end{pmatrix}. \quad (9)$$

3.2.2. Neural Networks

One notable model in machine learning is called **Neural Networks** (NN), which are inspired by the neurobiology of the brain, and hence the name. Mathematically, NNs utilize (repeated) composition of nonlinear transformations in developing their architecture. The definition is fairly simple (40); given a set of data inputs $\mathbf{x}_i \in \mathbb{R}^n$ and labels $\mathbf{y}_i \in \mathbb{R}$, $i = 1, \ldots, N$ as defined above, a NN is aimed to approximate a function of the form:

$$f_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = \sigma_L\left(\mathbf{w}^{(L)} \cdot \sigma_{(L-1)}\left(\mathbf{w}^{(L-1)} \cdot \cdots \sigma_1\left(\mathbf{w}^{(0)} \cdot \mathbf{x} + \mathbf{b}^{(0)}\right) + \mathbf{b}^{(L-2)}\right) + \mathbf{b}^{(L)}\right), \quad (10)$$

via adjusting unknown coefficients $\{\mathbf{w}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell+1}}\}_{\ell=0}^{L}$ and $\{\mathbf{b}^{(\ell)} \in \mathbb{R}^{n_\ell}\}_{\ell=0}^{L}$ such that the loss function is minimal between the data and the NN model:

$$\mathcal{L}\left(\left\{\mathbf{w}^{(\ell)}\right\}, \left\{\mathbf{b}^{(\ell)}\right\}\right) = \sum_{i=1}^{N} g\left(\mathbf{y}_i, f_{\mathbf{w},\mathbf{b}}\left(\mathbf{x}_i\right)\right), \quad (11)$$

where in Equation (10), the given functions $\sigma_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_\ell}$ are called *activation functions*, which are fixed for a particular architecture. The integer L of max composition is interpreted as layers with index $\ell = 0, \ldots, L$ denoting the layer number as shown in **Figure 5A** and n_ℓ is an integer denotes the number of **nodes (neurons)** in layer ℓ. The function g in Equation (11) should also be fixed depending on data query type. For continuous labels $\mathbf{y}_i$, such NN is called a regression prediction function with $g(\mathbf{y}, \mathbf{h}(\mathbf{x})) = ||\mathbf{y} - h(\mathbf{x})||^2$ typically adopted for an arbitrary loss function $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$. For discretized labels of multidimensions $\mathbf{y} = (y_1, \ldots, y_m)$, such NN is called a classification prediction function with cross entropy loss function $g(\mathbf{y}, \mathbf{h}(\mathbf{x})) = \sum_{k=1}^{m} [y_k \log(h_k(\mathbf{x})) + (1 - y_k) \log(1 - h_k(\mathbf{x}))]$ typically chosen with $\mathbf{h} = (h_1, \ldots, h_m)$.

In practice, there are several choices for activation functions σ_i, such as sigmoid, ReLu, eLu, Leaky ReLU function, etc., whose effectiveness usually depends on the nature of the dataset and the problem in question. The terms relating *forward dynamics*, *error backward propagation*, and *weights gradient descent* are technical procedures for estimating the unknown coefficients $\{\mathbf{w}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell+1}}\}_{\ell=0}^{L}$ and $\{\mathbf{b}^{(\ell)} \in \mathbb{R}^{n_\ell}\}_{\ell=0}^{L}$ from Equation (11). Although the design construction of an NN is relatively simple, the proper optimization of its parameters could be tedious numerically (40, 41).

In general, it is conventionally dubbed a *deep neural network* (DNN) when the number of hidden layers exceed 2, or $L \geq 3$. These neural networks are widely applied and are the foundations for the emerging field of *deep learning*, which is currently overperforming many of the classical machine learning techniques.

3.2.3. Deep learning Models

In KBR-ART, one expects that the processes involved in outcome modeling and adaptation procedures can be quite complex in nature for individualizing patient's treatment according to her/his predicted response over the course of fractionated therapy. There are few advanced data-driven models, mostly deep learning based, which can effectively into consideration such temporal information for updating knowledge and interactions between physical and biological variables for adapting therapy. In the following, we will briefly describe some of the main deep learning technologies in the literature.

3.2.3.1. Convolutional Neural Networks (CNNs)

CNNs are best known for image recognition and image-related prediction. The idea of CNN stemmed from the successful application of the signal processing operation of *convolution* in imaging processing, which was then been applied into neural networks for handling image related tasks. A CNN typically consists of several *convolutional layers, pooling layers, with activation functions* (42), where the convolution layer is the core component that applies an efficient convolutional filter (kernel) to the data in contrast to the tedious matrix operations described earlier with standard NN. In the case of a 2D image of size $L_1 \times L_2$ with multi color channels (C_1), the data are represented by a 3d-tensor $\mathcal{I} = \{\mathcal{I}_{i,j,\alpha}\}_{i=1,j=1,\alpha=1}^{L_1,L_2,C_1} \in \mathbb{R}^{2+1}$, a convolutional layer with stride s renders an output image (also called **feature maps**) $\widetilde{\mathcal{I}}$ (of size $\widetilde{L}_1 \times \widetilde{L}_2$ with C_2 channels) by applying the following convolution process (42).

$$\widetilde{\mathcal{I}}_{k,\ell,\beta} = \sum_{m,n,\alpha}^{L_1,L_2,C_1} w_{m,n,\alpha,\beta} \cdot \mathcal{I}_{s(k-1)+m,s(\ell-1)+n,\alpha} \quad \left(\text{image convolution. } k = 1, \ldots, \widetilde{L}_1, \ell = 1, \ldots, \widetilde{L}_2, \beta = 1, \ldots C_2\right), \quad (12)$$

here, $\mathbf{w} = \{w_{m,n,\alpha,\beta}\}_{m=1,n=1,\alpha=1,\beta=1}^{L_1,L_2,C_1,C_2} \in \mathbb{R}^{2+1+1}$ is a 4-tensor convolutional kernel. Such convolution process with stride is then equivalent to a regular convolution with image downsampling procedure. In fact, one can recognize that CNNs use these kernels in a neural network to "capture" local information within a neighborhood while "blocking" distant information or less related ones, as depicted in **Figure 5B**. Activation functions in CNNs have similar choices as a standard NN, Equation (10), mentioned above. CNNs has been successfully applied for image segmentation (43–46) in radiotherapy and for modeling of rectal toxicity in cervical cancer using transfer learning (47, 48). This will be further discussed in Section 3.3.1.

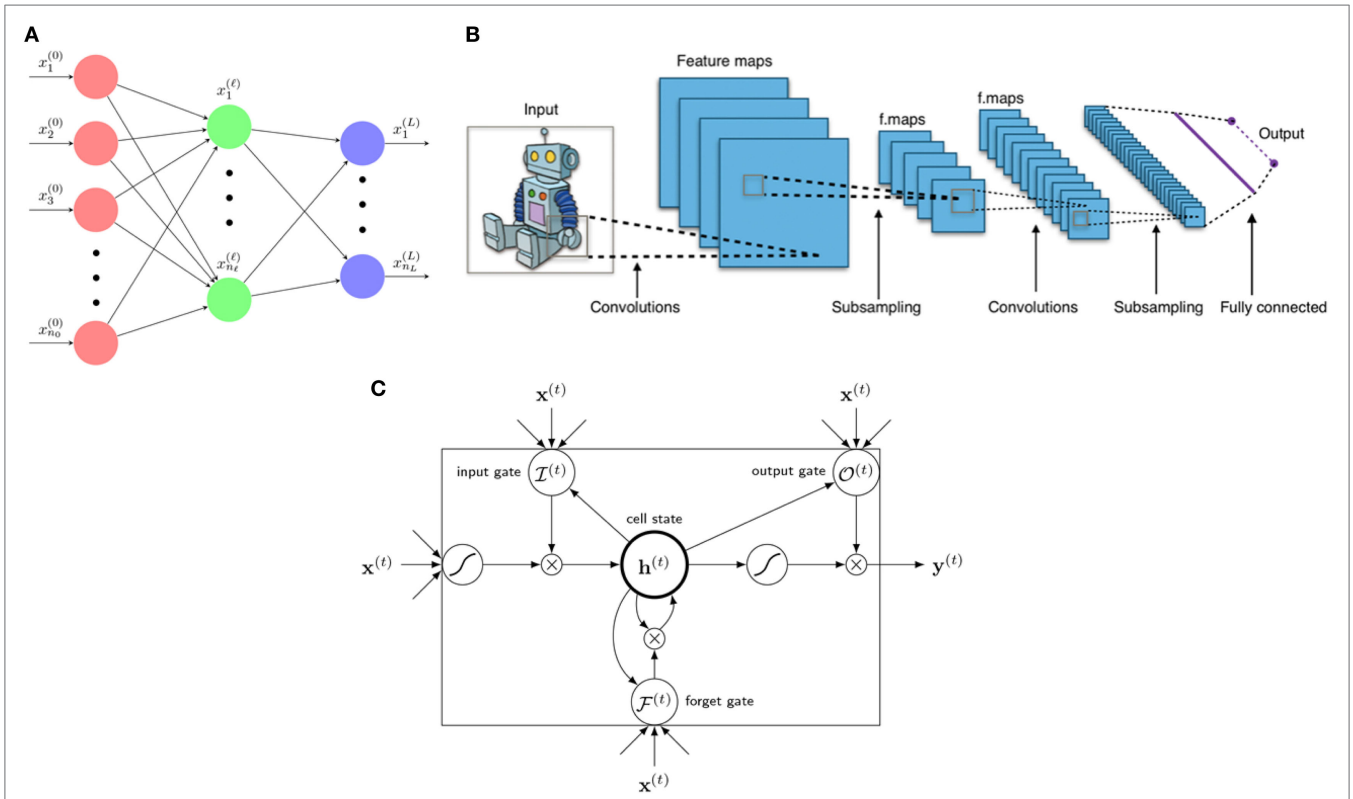

FIGURE 5 | Three main architectures of deep learning. **(A)** A neural network is a composition function interpreted to have several layers from $0, \ldots, L$ with neural nodes $(n_0, \ldots, n_\ell, \ldots, n_L)$. Variables $\mathbf{x}^{(\ell)} = (x_1^{(\ell)}, \ldots, x_{n_\ell}^{(\ell)})$ are called neurons of layer ℓ. **(B)** A typical architecture of a CNN consists of several layers including linear convolutional layers, pooling layers, and a final fully connected layer for classification (or regression). A kernel (filter) acts as a mask operating only on neighboring information (pixels) yet blocking distant information. [Figure created by Aphex34 distributed under a CC BY-SA 4.0 license (from Wikimedia Commons)]. **(C)** An LSTM unit consists of 1 cellstate $\mathbf{h}^{(t)}$ and 3 gates: forget gate $\mathcal{F}^{(t)}$, input gate $\mathcal{I}^{(t)}$, and output gate $\mathcal{O}^{(t)}$, with $\mathbf{x}^{(t)}$ as input and $\mathbf{y}^{(t)}$ as final (prediction) output.

3.2.3.2. Recurrent Neural Networks (RNNs)

RNNs are another variant of neural networks especially useful for learning sequential data, such as voice, text data, and handwriting. Therefore, it is also considered ideal for sequential adaptive radiotherapy with changing dose fractionations. In this case, suppose that we have sequential data $\{\mathbf{x}^{(t)} \in \mathbb{R}^n | \, t \in T\}$ as an input and $\{\tilde{\mathbf{y}}^{(t)} \in \mathbb{R}^m | \, t \in T\}$ as the corresponding labels where T denotes an index set (continuous or discrete) labeling separation across time steps. An important property of a RNN is that it introduces hidden units $\{\mathbf{h}^{(t)} \in \mathbb{R}^k | \, t \in T\}$ for making neural network deeper in increasing sequential prediction. A RNN is then aimed to learn the relationships between data $\{\mathbf{x}^{(t)} \in \mathbb{R}^n\}$ and labels $\{\tilde{\mathbf{y}}^{(t)}\}$ via hidden units $\{\mathbf{h}^{(t)} \in \mathbb{R}^k\}$ dynamically.

An RNN is designed to model the hidden variables via the recursive function $f_\theta : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}^k$.

$$\mathbf{h}^{(t)} = f_\theta \left(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)} \right) \in \mathbb{R}^k, \tag{13}$$

where θ usually serves as unknown neural weights to be solved, as $\{\mathbf{w}^{(\ell)}, \mathbf{b}^{(\ell)}\}_{\ell=0}^{L}$ in Equation (10).

One of the most successful RNN is the Long Short-Term Memory (LSTM). An LSTM is a state-of-the-art RNN model effective in sequential learning utilizing the so-called *gated units*, who learns by itself to store and forget internal memories when needed such that it is capable of creating long-term dependencies and paths through time, **Figure 5C**. A LSTM is constructed by 3 gates and 1 cell (hidden) state built up by the following equations.

$$\begin{aligned}
\mathcal{F}^{(t)} &= \sigma_g + (W_\mathcal{F} \cdot \mathbf{x}^{(t)} + U_\mathcal{F} \cdot \mathbf{h}^{(t-1)} + \mathbf{b}_\mathcal{F}) \in [0,1] \\
\mathcal{I}^{(t)} &= \sigma_g + (W_\mathcal{I} \cdot \mathbf{x}^{(t)} + U_\mathcal{I} \cdot \mathbf{h}^{(t-1)} + \mathbf{b}_\mathcal{I}) \in [0,1] \\
\mathcal{O}^{(t)} &= \sigma_g + (W_\mathcal{O} \cdot \mathbf{x}^{(t)} + U_\mathcal{O} \cdot \mathbf{h}^{(t-1)} + \mathbf{b}_\mathcal{O}) \in [0,1] \\
\mathbf{h}^{(t)} &= \mathcal{F}^{(t)} \circ \mathbf{h}^{(t-1)} + \mathcal{I}^{(t)} \circ \sigma_\mathbf{h} \left(W_\mathbf{h} \cdot \mathbf{x}^{(t)} + U_\mathbf{h} \cdot \mathbf{h}^{(t-1)} + \mathbf{b}_\mathbf{h} \right) \\
\mathbf{y}^{(t)} &= \mathcal{O}^{(t)} \circ \sigma_\mathbf{y}(\mathbf{h}^{(t)}),
\end{aligned} \tag{14}$$

where $\sigma_g, \sigma_\mathbf{h}, \sigma_\mathbf{y}$ are 3 non-linear activation functions depending on one's choice, $\{\mathcal{F}^{(t)}, \mathcal{I}^{(t)}, \mathcal{O}^{(t)}\}$ are called the **forget gate**, **input gate**, and **output gate** at time t, respectively.

The 3 gates, with all their numerical values in [0,1], are used to control and determine when and how much should the previous information be kept or forgotten. The unknown parameters of an LSTM are $(W_h, U_h, \mathbf{b}_h)$ and $\{(W_\alpha, U_\alpha, \mathbf{b}_\alpha) \, | \, \alpha = \mathcal{F}, \mathcal{I}, \mathcal{O}\}$ and, therefore, an LSTM unit generally possesses four times parameters than a plain neural net in Equation (10) requiring a large amount of data for training. RNNs have been evaluated in radiotherapy for respiratory motion management (49). An interesting approach combining RNN with CNN was used for pancreas segmentation

on both CT and MRI datasets, which mitigated the problem of using spatial smoothness consistency constraints (50, 51).

The previously presented machine learning methods do not allow visualization of the system dynamics and act primarily as a black box mapping from the input to the output data and are referred to as *discriminant* models. Alternatively, system dynamics of mapping input to output data can be revealed using so-called *generative* models. A common example of such models is Bayesian networks, which will be discussed next.

3.2.4. Bayesian Networks

Bayesian networks (BNs) are a class of probabilistic graphical models (GM) corresponding to directed acyclic graphs (DAGs), which are also named as belief networks. BNs combine graph theory, probability theory, computer science, and statistics to represent knowledge in an uncertain domain. They are popular in the societies of statistics, machine learning, and artificial intelligence. Especially, BNs are mathematically rigorous and intuitively understandable, which enable an effective way to represent and compute the joint probability distribution (JPD) over a set of random variables (52).

Each BN includes the sets of nodes and directed edges. While the former indicate random variables represented by circles, the latter display direct dependencies among these variables illustrated by arrows between nodes. In a BN, an arrow from node X_i to node X_j shows a statistical dependence between them, which indicates that a value of variable X_j depends on that of variable X_i, or variable X_i "affects" X_j. Also, their relationship can be described as follows: Node X_i is a parent of X_j and node X_j is the child of X_i. In general, the set of nodes that can be reached on a direct path from the node is named as the set of its descendants, and the set of nodes from which the node can be reached on a direct path is called as the set of its ancestor nodes (53).

The DAG structure guarantees that no node can be its own ancestor or its own descendant, which is of vital importance to the factorization of the JPD of a collection of nodes. A BN is designed to reflect a conditional independence statement, where each variable is independent of its nondescendants in the BN given its parents. This property is used to significantly reduce the number of parameters required to characterize the JPD of the variables. Especially, this reduction leads to an efficient way in computing the posterior probabilities given the evidence (52, 54, 55).

Moreover, the parameters of the BN are described in a manner following a Markovian property, where the conditional probability distribution (CPD) of each node only depends on its parents. These conditional probabilities are often represented by a table for discrete random variables to list the conditional probability that a child node takes on each of the feasible values from each combination of values of its parents. The joint distribution of a collection of variables can be obtained uniquely by these conditional probability tables (CPTs).

Generally, a BN B can be considered as a DAG that represents a joint probability density function over a set of random variables V. The BN is defined by a pair $\mathrm{B} = \langle G, \phi \rangle$, where G is the DAG whose nodes $X_1, X_2, \ldots, X_n$ denotes random variables, and whose edges indicate the direct dependencies between them. The graph G includes independence assumptions, where each variable X_i is independent of its nondescendants given its parents in G. The second component ϕ represents the set of parameters of the BN. This set contains the parameter $\theta(x_i|\pi_i) = P_B(x_i|\pi_i)$ for each realization x_i of X_i conditioned on π_i, which is the set of parents of X_i in G. Then, B describes a unique JPD over V:

$$P_B(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i|\pi_i) = \prod_{i=1}^{n} \theta_{X_i|\pi_i}, \qquad (15)$$

where if X_i does not have parents, its probability distribution is considered to be unconditional; otherwise it is conditional. Once the variable indicated by a node is observed, the node is considered as an evidence node; otherwise the node is treated as a hidden or latent node. Because of their generative nature, BNs have been widely applied for modeling radiotherapy errors (56, 57) and outcomes (58–62). This will be further discussed in Section 3.3.2.

3.3. Example Application of Machine Learning to Outcome Modeling

As examples of application of modern machine learning to outcome modeling, in the following, we discuss application of a discriminant modeling approach by CNN of rectal toxicity and a generative modeling approach by BN for lung toxicity.

3.3.1. NTCP Modeling of Rectal Toxicity Using CNN

Zhen et al. (47) studied the possibility of modeling rectal toxicity in cervical cancer using CNNs from unfolded rectum surface dose maps (RSDMs) (63) with the help of transfer learning, as depicted in **Figure 6**. A retrospective data of 42 cervical cancer patients were studied. These patients were treated with external beam radiotherapy (EBRT) and/or brachytherapy (BT). The EBRT was delivered in 25 fractions (2 Gy/fraction) and BT was delivered in 4–6 fractions (6–7 Gy/frac).

For transfer learning, CNN of VGG-16 (64) was chosen as optimal architecture, which consists of 16 convolutional layers of suitable sizes including up to 138 million parameters. The VGG-16 is pretrained using a publicly annotated natural images database (ImageNet). The finetuned VGG-16 on the cervix cancer dataset with ADASYN method for imbalance correction, achieved an AUC of 0.89 on leave-one-out cross validation for rectal toxicity prediction. In addition to a successful model building of relating RSDMs to toxicity, Zhen et al. also attempted to interpret what and how CNNs "view" an RSDM, where the method of Grad-CAM map (65) was utilized to unveil the nature of the CNN learnt features (**Figure 7**). From **Figure 8**, one finds that the Grad-CAM interpreted maps (d, e) (from mapping CNN weights) have high consistency of distinct image patterns with toxicity (b) and non-toxicity (c) that were recognizable by human eyes. Therefore, by visualizing the CNN model, one can have better understanding of the features learned by the machine learning algorithm.

3.3.2. NTCP Modeling of Lung Toxicity Using Bayesian Networks

Radiation pneumonitis of grade 2 or above (RP2) is a major radiation-induced toxicity in NSCLC radiotherapy, and it may depend on radiation dose, the patients' clinical, biological, and genomic characteristics. In order to find appropriate treatment

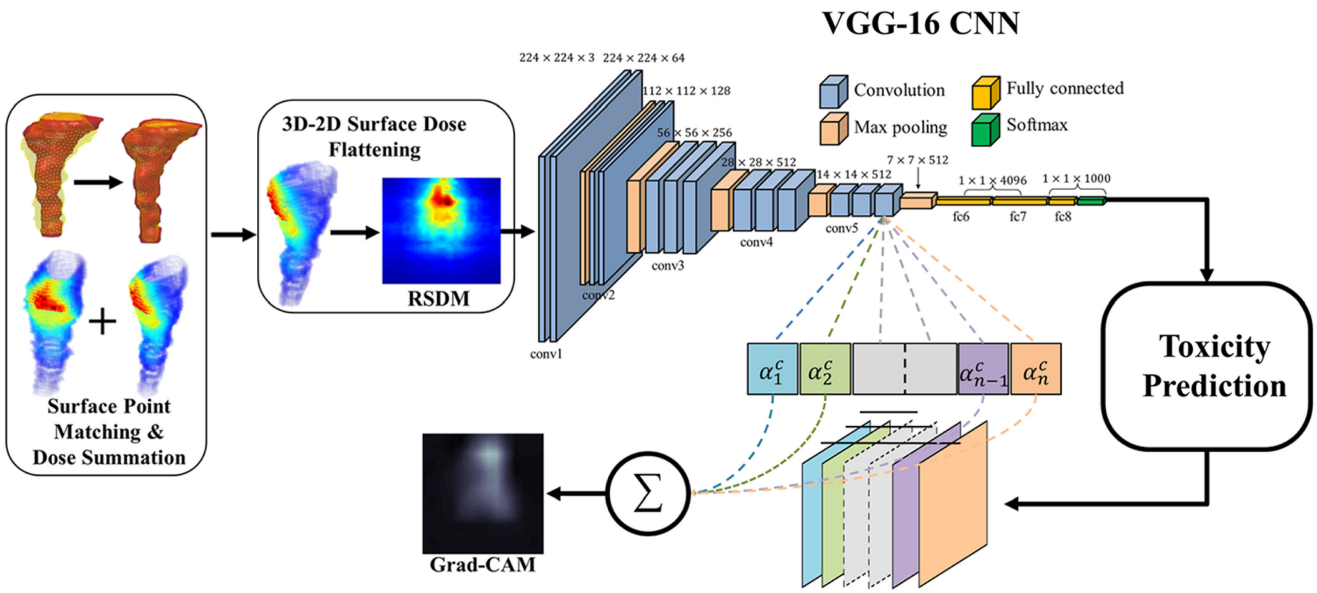

FIGURE 6 | The workflow of the rectum toxicity study in (47) using VGG-16 receiving 2D RSDM image input with Grad-CAM map as interpretation of CNN weights. [© Institute of Physics and Engineering in Medicine. Reproduced by permission of IOP Publishing. All rights reserved.]

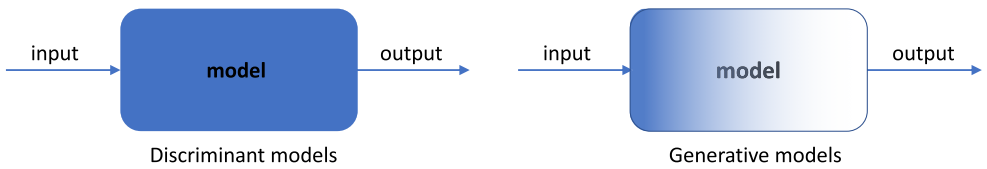

FIGURE 7 | Comparison of discriminant and generative models, where the gradient coloring on generative models indicates the model transparency.

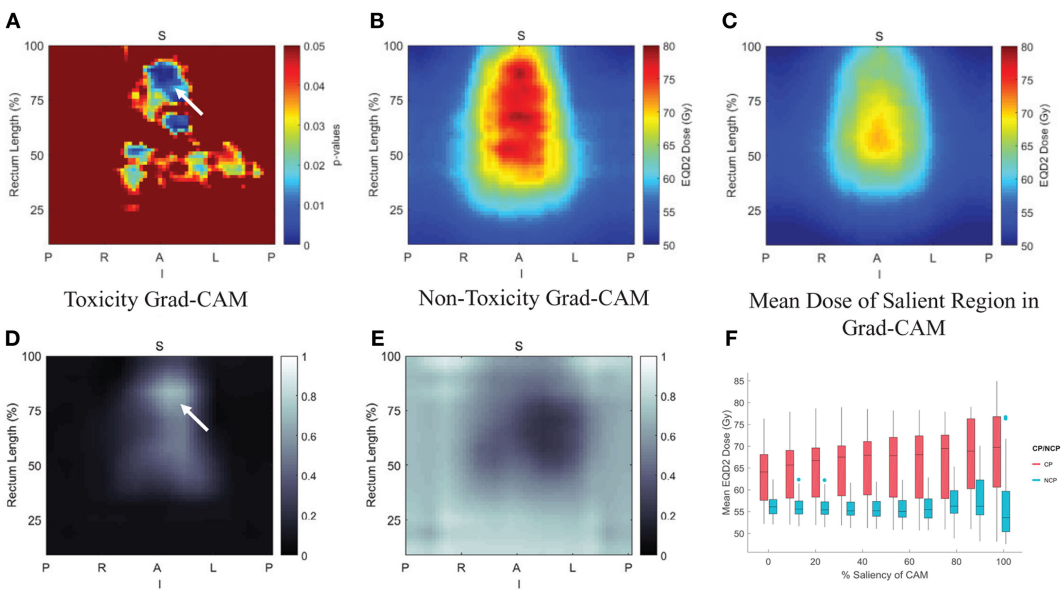

FIGURE 8 | Pixelwise *p*-value map were shown in **(A)** with small $p < 0.05$, **(B,C)** are the average rectum RSDM of the toxicity and non-toxicity patients; and **(D,E)** are average Grad-CAM map of the toxicity and non-toxicity groups. **(F)** Box plot of the mean dose in different salient regions extracted from the Grad-CAM map. Details see (47).

plans and improve patients' therapeutic satisfaction, a systematic machine learning approach needs to be developed to find the most important features from the high dimensional dataset and to discover the relationships between them and RP2 for clinical decision-making. Thus, a BN approach was developed to explore interpretable biophysical signaling pathways influencing RP2 from a heterogeneous dataset including single nucleotide polymorphisms (SNPs), micro RNAs (miRNAs), cytokines, clinical data, and radiation treatment plans before and during the course of radiotherapy of NSCLC patients.

In this BN implementation, the dataset described in Section 2.5.1 with 79 patients (21 cases of RP2) was used for model building and 46 additional patients were reserved for independent model testing. The BN approach mainly included a large-scale Markov blanket (MB) method to select relevant predictors, and a structure learning algorithm to find the optimal BN structure based on Tabu search and the performance evaluation of outcome prediction (24). *K*-fold cross-validation was used to guard against over-fitting, and the area under the receiver-operating characteristics (AUC) curve was utilized as a prediction metric.

The large-scale MB method intends to identify the most relevant variables of RP2 before or during the course of radiotherapy. **Figure 9A** shows the extended MB neighborhoods of RP2 before radiation treatment, where the MB of RP2 based on pretreatment training data is formed from "Mean_Lung_Dose," "pre_MCP_1," "pre_TGF_alpha," and "pre_eotaxin." In the meantime, each of these variables has its own MB neighborhood as shown in **Figure 9A**. For example, "V20," "nos3_Rs1799983," "stage," and "RP2" form the MB of "Mean_Lung_Dose." In this study, potential variables of the BN were identified from the extended MB neighborhoods within two layers of RP2. **Figure 9B** indicates the updated extended MB neighborhoods in an extended model after incorporating the slopes of cytokine levels before and during-treatment (SLP) as the patients' responses during the radiation treatment. Although the MB of RP2 during the radiation treatment based on the whole training dataset keeps the same as that in **Figure 9A**, the MB of "Mean_Lung_Dose" has been updated, and it includes patients' cytokine responses such as "SLP_IL_17," "SLP_GM_CSF." **Figures 9C,D** illustrate biophysical signaling pathways from the patients' relevant variables to RP2 risk based on pretreatment and during BN model building, respectively. The results of internal cross-validation show that the performance of the BN yielded an AUC = 0.82, and it was improved by incorporating during treatment cytokine changes to AUC = 0.87. In the testing dataset, the pre- and during AUCs were 0.78 and 0.82, respectively. It turns out that the BN approach allows for

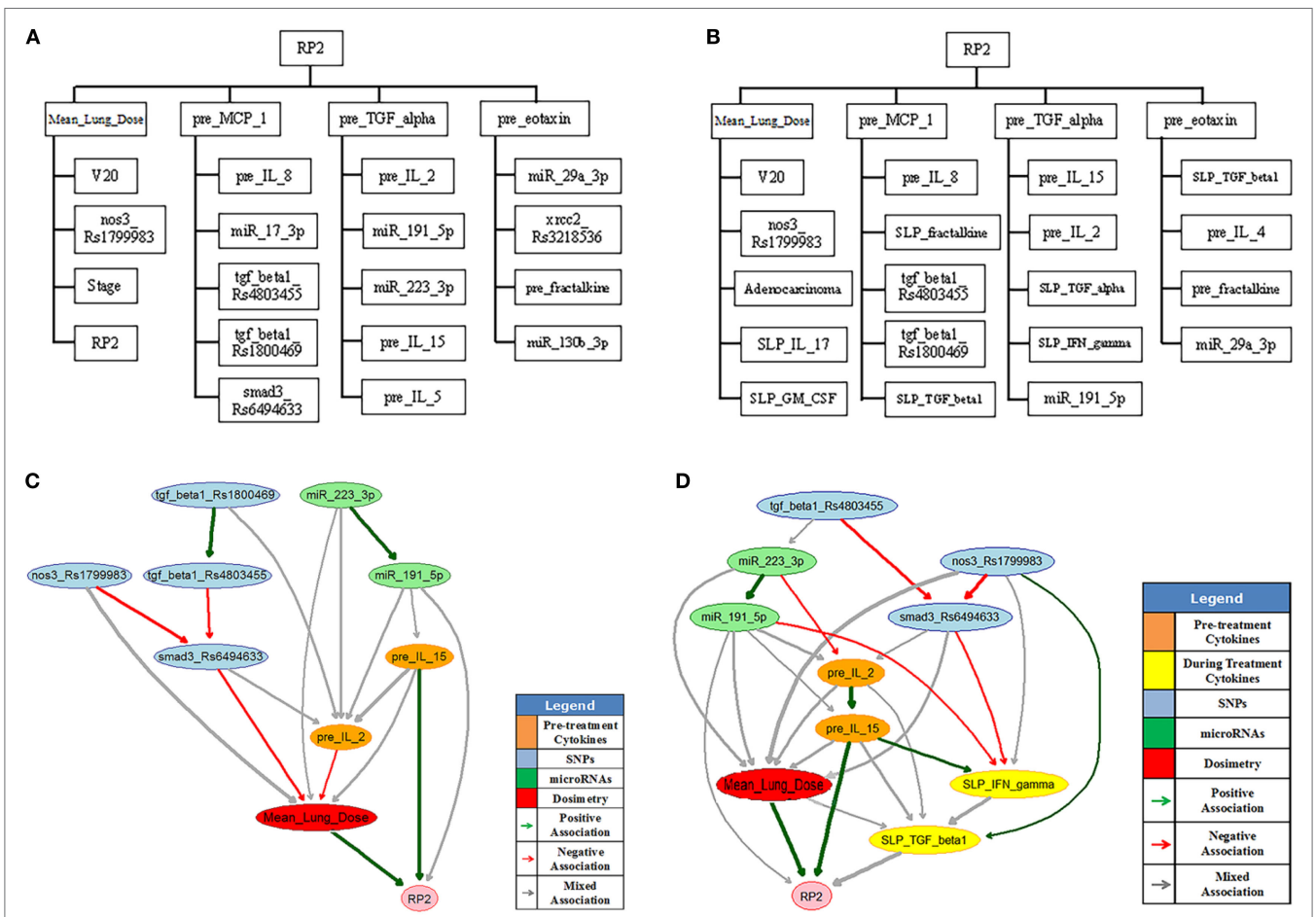

FIGURE 9 | The extended MB neighborhoods of RP2 before **(A)** and during **(B)** radiation treatment, where the upper level shows the inner family of RP2 and the lower levels show the next-of-kin for each of its member. Pretreatment BN **(C)** and during-treatment BN **(D)** for RP2 prediction [figures reprinted with permission].

unraveling of relevant biophysical features that lead to RP2 risk and prediction of RP2, and this prediction improved by incorporating during treatment information (24).

4. Q3: HOW TO ADAPT PLANS IN KBR-ART?

The precise estimation of treatment outcome is necessary step before deciding on the right course of action, since we desire to evaluate potential outcomes effects beforehand as we weigh the different alternatives for the best possible strategy (i.e., set of actions) to optimize the individual's treatment response. This is in a simplistic sense no different than playing board games or chess when a player may evaluate a dozen of options before carrying out a move. Therefore, by assuming one can attain accurate prediction estimates of TCP and NTCP, as discussed in the previous section, then, the final question to address in the context of KBR-ART is how to optimally adapt the plan (e.g., increase the tumor fraction dose) to achieve improved outcomes.

A utility function is usually required to estimate the total effect of a treatment plan weighting on both positive outcomes and the possible side effects caused. In RT, an example utility function called *complication-free tumor control (P^+)* can be used. The P^+ measures the performance of a treatment at each stage based on combined TCP and NTCP under the form $P^+ = U(\text{TCP, NTCP}; \theta)$ where P^+ indicates *probability of a positive treatment outcome*. One linear form is particularly simple and effective (66) where:

$$P^+ = \text{TCP} \times (1 - \text{NTCP}) \tag{16}$$

Notably, some other functional forms may be used as well, such as Equation (42).

In the practice of KBR-ART, if one has already synthesized relevant knowledge (clinical, dosimetric data, . . . etc.) from Section 2 with variables $x_1, \ldots, x_n$ as predictors and applied analytical/data-driven models in Section 3, then we can derive models of TCP and NTCP in the form $\text{TCP} = f_{\text{TCP}}(x_1, \ldots, x_n)$ and $\text{NTCP} = f_{\text{NTCP}}(x_1, \ldots, x_n)$ based on retrospective data such that the P^+ response estimation function reads:

$$P^+ = U\left(f_{\text{TCP}}(x_1, \ldots, x_n), f_{\text{NTCP}}(x_1, \ldots, x_n); \theta\right) \tag{17}$$

With the response estimation defined by the P^+ utility functions, next, we design a scheme for treatment adaptation. Machine learning based on reinforcement learning (RL) is a suitable approach for realizing plan adaptation as it can search over all possible decisions to maximize the P^+ function as *rewards* and identify the best *policy* (e.g., dose per fraction) for the treatment planning.

4.1. Generalized KBR-ART Framework

The KBR-ART can be described by the following general formulation:

$$\begin{aligned} &\{\mathbf{x}^{(t)} \in \mathbb{R}^n | t \in T\}, \quad \{\mathbf{y}^{(t)} \in \mathbb{R}^m | t \in T\}, \quad \{\mathbf{u}^{(t)} \in \mathbb{R}^p | t \in T\} \\ &\mathcal{L}(\{\mathbf{x}^{(t)}\}, \{\mathbf{y}^{(t)}\}, \{\mathbf{u}^{(t)}\}; \theta), \quad \mathcal{C}(\{\mathbf{x}^{(t)}\}, \{\mathbf{y}^{(t)}\}, \{\mathbf{u}^{(t)}\}; \phi), \end{aligned} \tag{18}$$

where $\mathbf{x}^{(t)}$ is the state of a system at time $t \in T$, $\mathbf{y}^{(t)}$ is the observation of state $\mathbf{x}^{(t)}$, $\mathbf{u}^{(t)}$ is the controls for the system to influence next states $\mathbf{x}^{(t+1)}$, and $\mathcal{L}(\{\mathbf{x}^{(t)}\}, \{\mathbf{y}^{(t)}\}, \{\mathbf{u}^{(t)}\}; \theta)$ is a loss function serving a specific purpose for the system to be minimized over temporal information $\mathbf{x}^{(t)}$, $\mathbf{y}^{(t)}$, and $\mathbf{u}^{(t)}$ along with some constraints $\mathcal{C}(\{\mathbf{x}^{(t)}\}, \{\mathbf{y}^{(t)}\}, \{\mathbf{u}^{(t)}\}; \phi)$. Any of the vectors $\mathbf{x}^{(t)}$, $\mathbf{y}^{(t)}$, and $\mathbf{u}^{(t)}$ can be real-valued vectors or vectors of random variables such that the temporal sequences can be deterministic or a random processes adaptation. Although dimensions of n, m, p may be infinite in Equation (18), almost all real-life implementations are finite dimensions. Equation (18) may apply to many legacy ART approaches in different manners. In the following, we provide a brief overview for alternative ART approaches.

4.1.1. Linear Feedback ARTs

Traditionally, linear feedback (loop) control systems are considered as viable implementations of ART, where most of the adaptive feedback is based on imaging information such as CT and/or MRI. Generally, there are two types of control systems: *open-loop* and *closed-loop*.

With notations in Equation (18), a *linear loop control* is generally described by two sets of linear equations:

$$\dot{\mathbf{x}}^{(t)} = A\mathbf{x}^{(t)} + B\mathbf{u}^{(t)}, \qquad \mathbf{y}^{(t)} = C\mathbf{x}^{(t)} \tag{19}$$

$$\dot{\tilde{\mathbf{x}}}^{(t)} = A\tilde{\mathbf{x}}^{(t)} + B\mathbf{u}^{(t)} + L\delta\mathbf{y}^{(t)}, \qquad \tilde{\mathbf{y}}^{(t)} = C\tilde{\mathbf{x}}^{(t)}, \tag{20}$$

where in Equation (19), A, B, L, C are linear operators, $\mathbf{y}^{(t)}$ is the *observation* of the actual state $\mathbf{x}^{(t)}$, and $\{u^{(t)} | t \in \mathcal{T}\}$ represents controls of the system as adaptations for the treatment of a radiotherapy. Equation (20) as a similar copy of Equation (19) describes the **estimation** $\tilde{\mathbf{x}}^{(t)}$, $\tilde{\mathbf{y}}^{(t)}$ of the corresponding variables $\mathbf{x}^{(t)}$, $\mathbf{y}^{(t)}$ of the system, with $\delta\mathbf{y}^{(t)} \overset{def}{=} \mathbf{y}^{(t)} - \tilde{\mathbf{y}}^{(t)}$ as the **estimation error** of the observed state and in turn shall be used as the **feedback** in the subsequent iterations. With $L \neq 0$, the system constantly receiving the estimation error shall adjust itself accordingly, and thus such is called a **closed-loop** control system, **Figure 10A**.

Incidentally, in the perfect case, the three characters $\mathbf{x}^{(t)}$, $\tilde{\mathbf{x}}^{(t)}$, $\tilde{\mathbf{y}}^{(t)}$ shall coincide into one with $C = I$, $\delta\mathbf{y}^{(t)} = 0$ and thus the Equations (19) and (20) reduce to one. However, in most of cases, they tend to split. In a system, where the matrix L vanishes, it becomes an **open-loop** control system since any feedback signal $\delta\mathbf{y}^{(t)}$ from the system is not considered, **Figure 10B**. An obvious drawback of the open-loop system is the estimation instability, which can be easily seen from Equations (19) and (20) as the quantity $\delta\mathbf{x}^{(t)} \overset{def}{=} \mathbf{x}^{(t)} - \tilde{\mathbf{x}}^{(t)}$ describing the **estimation error** is subject to the state equation $d/dt(\delta\mathbf{x}^{(t)}) = A \cdot \delta\mathbf{x}^{(t)}$ with $L \equiv 0$. The solution $\delta\mathbf{x}^{(t)} = e^{AT} \cdot \delta\mathbf{x}^{(0)}$ indicates that the error has exponential growth as time elapses such that soon an open-loop system easily becomes unreliable. On the other hand, by receiving a feedback signal due to a close-loop system ($L \neq 0$) can improve reliability, as the evolution $\delta\mathbf{x}^{(t)} = e^{(A-LC)t} \cdot \delta\mathbf{x}^{(t)}$ will converge by suitable choice of a **gain** L such that the eigenvalues $|\lambda_i (A - LC)| < 1$. In a linear control problem, the control is modeled by $\mathbf{u}^{(t)} = -K\mathbf{x}^{(t)}$ with a constant matrix K such that Equation (20) reads:

$$\dot{\tilde{\mathbf{x}}}^{(t)} = (A - LC)\tilde{\mathbf{x}}^{(t)} - BK\,\mathbf{u}^{(t)} \tag{21}$$

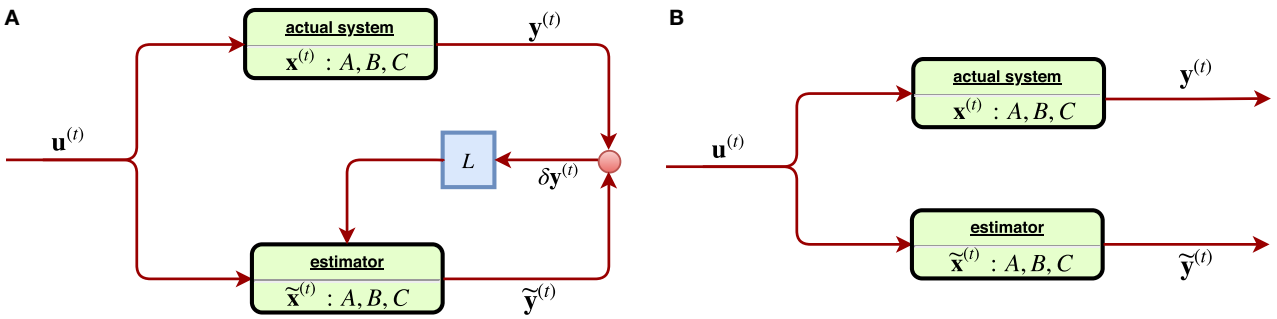

FIGURE 10 | **(A)** [Left] a linear closed-loop control system, where the feedback signal of **estimation error** of the observed state $\delta\mathbf{y}^{(t)}$ is received through a gain (matrix) L. **(B)** [Right] a linear open-loop control system where the feedback signal $\delta\mathbf{y}^{(t)}$ was not considered. Such system tends to suffer from estimation error instability, $\lim_{t\to\infty} \|\delta\mathbf{x}^{(t)}\| \to \infty$.

In control theory, one may also consider a loop-control system with small uncertainty. Typically, by considering stochasticity, the system can become more stable and robust. The (time) discretized linear control with a random process starts with extension of Equation (19) as:

$$\begin{aligned}\mathbf{x}^{(t)} &= A\cdot\mathbf{x}^{(t-1)} + B\cdot\mathbf{u}^{(t)} + \mathbf{w}^{(t)}\\ \mathbf{y}^{(t)} &= C\cdot\mathbf{x}^{(t-1)} + \mathbf{v}^{(t)},\end{aligned} \tag{22}$$

where the two random processes $\{\mathbf{w}^{(t)}\}$ and $\{\mathbf{v}^{(t)}\}$ denote the noise of state $\mathbf{x}^{(t)}$ and observation $\mathbf{y}^{(t)}$ assumed multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{Q}^{(t)})$ and $\mathcal{N}(\mathbf{0}, \mathbf{R}^{(t)})$, respectively. Kalman filters are then a common analysis for deriving optimal estimation of $\delta\mathbf{x}^{(t)}$. In (67), Keller et al. established a linear stochastic closed-loop system that utilized Kalman filters (68) to derive optimal control law. They assumed an image-guided radiotherapy, which attempts to provide optimal correction strategies for setup errors, which can also take the measurement uncertainties into account. Let $\mathbf{x}^{(t)} = \mathbf{x}_1^{(t)} + \mathbf{x}_2^{(t)} \in \mathbb{R}^3$ denote the difference between the actual and planned positions of the center-of-mass of the clinical tumor volume (CTV), i.e., the daily displacement $\mathbf{x}^{(t)}$ containing (1) the setup error $\mathbf{x}_1^{(t)}$ (displacement of bony structures) and (2) the organ motion (displacement $\mathbf{x}_2^{(t)}$ with respect to the bony structures). Decompose $\mathbf{x}^{(t)}$ into two parts $\mathbf{x}^{(t+1)} = \mathbf{u}^{(t)} + \mathbf{w}^{(t)}$ with $\mathbf{u}^{(t)} = \mathbf{u}_1^{(t)} + \mathbf{u}_2^{(t)}$ called the **systematic component** and $\mathbf{w}^{(t)} = \mathbf{w}_1^{(t)} + \mathbf{w}_2^{(t)}$ called the **random component**, where the subindex "1" and "2" refer to setup errors and organ motion, respectively. Together, they modeled the ART displacement with a stochastic linear system:

$$\begin{aligned}\mathbf{x}^{(t+1)} &= \mathbf{u}^{(t)} + \mathbf{w}^{(t)}\\ \mathbf{y}^{(t)} &= \mathbf{x}^{(t)} + \mathbf{v}^{(t)}\end{aligned} \tag{23}$$

where $\mathbf{y}^{(t)}$ is the observation of $\mathbf{x}^{(t)}$. By defining the estimation of state $\mathbf{x}^{(t)}$ as $\tilde{\mathbf{x}}^{(t)} \overset{def}{=} P(\mathbf{x}\,|\,\mathbf{y}^0, \ldots, \mathbf{y}^{(t-1)})$ based on previous observations $\mathbf{y}^0, \ldots, \mathbf{y}^{(t-1)}$ as in Equation (20), Kalman filters are able to provide an optimal estimation of $\tilde{\mathbf{x}}^{(t)}$ such that the estimation error $\tilde{\mathbf{x}}^{(t+1)} \overset{def}{=} \mathbf{x}^{(t+1)} - \tilde{\mathbf{x}}^{(t)}$ is minimal. Immediately, they derived

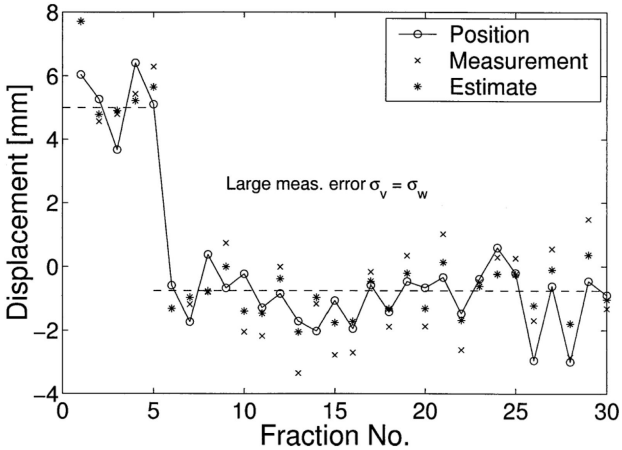

FIGURE 11 | A simulated displacements (circles) of 30 fractions was demonstrated in (67), where in the first 5 fractions, $u = 5$ is used. Their results showed that on average Kalman filters (asterisks) estimations $\tilde{\mathbf{x}}^{(t)}$ are closer to the (unknown) displacements than the measurements $\mathbf{y}^{(t)}$ (crosses).

the *optimal control law* $\mathbf{u}_{c*}^{(t)} = -\tilde{\mathbf{x}}^{(t)}$, which seems to be an intuitive result. A comparison was made with respect to an obvious control law that is "suboptimal" $\mathbf{u}_c^{(t)} = -\tilde{\mathbf{y}}^{(t)}$, which is merely the correction of observation itself. Subsequently, they attempted to measure the effectiveness of decisions given by Kalman filters $\mathbf{u}_{c*}$ and the observation $\mathbf{u}_c$ by computing

$$e \overset{def}{=} \frac{\sigma^2_{\mathbf{x}-\tilde{\mathbf{x}}}}{\sigma^2_{\mathbf{x}-\mathbf{y}}} \tag{24}$$

where $\sigma^2_{\mathbf{x}-\tilde{\mathbf{x}}}$ and $\sigma^2_{\mathbf{x}-\mathbf{y}}$ are two residue variances of different estimation toward the state $\mathbf{x}^{(t)}$. One simulated result was made to demonstrate the performance of Kalman filters in predictions of stochastic linear control system, as shown in **Figure 11** where a treatment of 30 fractions were simulated with the first 5 fractions, a random systematic error $u = +5$ mm and measurement noise $\sigma_v = \sigma_w = 1$ mm were imposed, which means the correction started only at the sixth fraction. Their results showed that on average Kalman filter estimations $\tilde{\mathbf{x}}^{(t)}$ are closer to the (unknown) displacements than the measurements $\mathbf{y}^{(t)}$, where in the first fraction the estimate equals the value of the measurement.

4.2. Nonlinear Feedback ARTs

It is natural to consider nonlinear feedback control for ARTs due to inherent complexity. In (69), Zerda et al. developed a nonlinear closed-loop ART for treatment planning. In particular, they proposed two algorithms: *Immediately Correcting Algorithm* and *Prudent Correcting Algorithm*. With the following notation corresponding to Equation (18),

$$\begin{aligned} \mathbf{x}^{(t)} = \mathbf{y}^{(t)} &\rightarrow \psi^{(t)} = \left(\psi^{(t)}_{\text{geometry}}, \psi^{(t)}_{\text{cumdose}}\right), \\ u^{(t)} = \xi^{(t)}(\{\mathbf{x}^{(t)}\}) &\rightarrow \beta^{(t)} = \xi^{(t)}\left(\psi^{(t)}\right), \\ \mathcal{L}(\{\mathbf{x}^{(t)}\}, \{\mathbf{y}^{(t)}\}, \{\mathbf{u}^{(t)}\}; \theta) &\rightarrow \sum_{v\in V} \alpha(v)\left(D_{\text{prescribed}}(v) - \psi^{|T|}_{\text{cumdose}}(v)\right)^2, \end{aligned} \tag{25}$$

where $v \in V$ is a voxel under consideration, $v \mapsto \alpha(v)$ is the importance factor, and the control is promoted as a nonlinear function of states, $\mathbf{u}^{(t)} = \xi^{(t)}(\{\mathbf{x}^{(t)}\})$ rather than the linear form $\mathbf{u}^{(t)} = -K \cdot \mathbf{x}^{(t-1)}$. With $\psi^{(t)}$ denoting the state of the ART system, it was assumed to consists of two parts: (1) **cumulative dose** $\psi^{(t)}_{\text{cumdose}}$ after $t \in T$ and (2) **patient's geometric model** obtained from conebeam CT (CBCT images) $\{\psi^{(t)}_{\text{geometry}} \,|\, t \in T\}$, where it was further assumed the geometry information interacts with the cumulative dose by the relation

$$\begin{aligned} \psi^{(t)}_{\text{cumdose}} &= \psi^{(t-1)}_{\text{cumdose}} \\ &+ D\left(v; \{\beta^{(t)} \,|\, t \in T\}, \{\epsilon^{(t)} \,|\, t \in T\}, \{\psi^{(t)}_{\text{geometry}} \,|\, t \in T\}\right) \end{aligned} \tag{26}$$

with the *dose delivery* function $D\left(v; \{\beta^{(t)}\}, \{\epsilon^{(t)}\}, \{\psi^{(t)}_{\text{geometry}}\}\right)$ related to delivery errors $\{\epsilon^{(t)}\}$, where it is always assumed vanishing throughout the paper (69). In other words, from Equations (25) and (26), the objective of the Immediately Correcting Algorithm is to minimize the following loss:

$$\begin{aligned} \mathcal{L}\left(\beta^1, \ldots, \beta^{|T|}\right) = \sum_{v\in V} \alpha(v) \Bigg(& D_{\text{prescribed}}(v) \\ & - \sum_{t\in T} D\left(v; \left\{\beta^{(t)}\right\}, \left\{\epsilon^{(t)}\right\}, \left\{\psi^{(t)}_{\text{geometry}}\right\}\right)\Bigg)^2 \end{aligned} \tag{27}$$

via an optimal sequence of dose fractionation (controls) $(\beta_1, \ldots, \beta_{|T|})$ to be found, and thus it is regarded as a special realization of the general scheme **Figures 12A,B**.

4.3. Stochastic ARTs

In (70), Bortfeld et al. developed a static *robust* optimization by treating the dose delivery problem of intensity modulated RT (IMRT) as a probabilistic problem with uncertainties. Using the notations in Equation (18) and letting $\mathbf{x}^{(t)}$ as a **breathing phase** (state) at time t, $\mathbf{u}^{(t)}$ as a **control probability function** over all breathing states, the **observed state** $\mathbf{y}^{(t)} = \mathbf{x}^{(t+1)}$:

$$\mathbf{x}^{(t)} \rightarrow x, \quad \mathbf{u}^{(t)} \rightarrow p(x), \quad \theta \rightarrow \{\Delta_{v,b,x}, w_b, \gamma, \theta_v\}, \tag{28}$$

we arrive at the loss function and constraints proposed by Bortfeld et al.

$$\begin{aligned} \text{minimize} \quad & \mathcal{L} = \sum_{v\in V}\sum_{x\in X}\sum_{b\in B} \Delta_{v,b,x}\, p(x) w_b \\ \text{subject to} \quad & \mathcal{C}_1 = \sum_{v\in V}\sum_{x\in X}\sum_{b\in B} \Delta_{v,b,x}\, \tilde{p}(x) w_b \geq \theta_v, \ \forall v \in \mathcal{T}, \ \tilde{p} \in P_U \\ & \mathcal{C}_2 = \sum_{v\in V}\sum_{x\in X}\sum_{b\in B} \Delta_{v,b,x}\, \tilde{p}(x) w_b \leq \gamma\theta_v, \ \forall v \in \mathcal{T}, \ \tilde{p} \in P_U. \end{aligned} \tag{29}$$

Essentially, they considered the dose (to be delivered) as an expectation value following a predefined probability distribution (PDF) over all breathing phases, $D_{v,b} = \mathbb{E}_x[\Delta_{v,b,x}] = \sum_{x\in X} \Delta_{v,b,x}\, p(x)$, where $v \in V$ denotes a voxel, $b \in B$ denotes

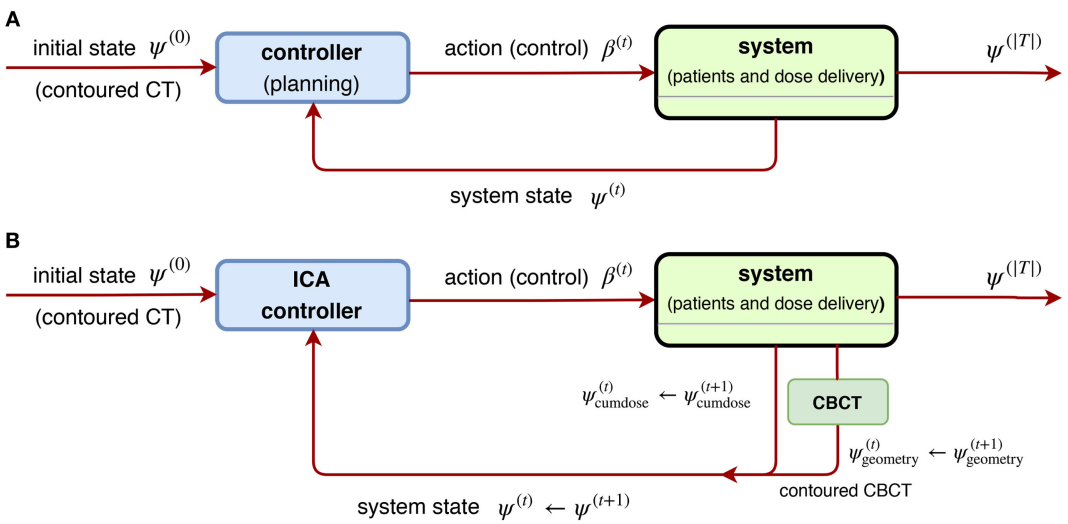

FIGURE 12 | **(A)** [Left] a general scheme of a non-linear closed-loop feedback control proposed by Zerda et al. (69), where a system feedback $\psi^{(t)}$ was received after fraction $t \in T$ is completed. **(B)** [Right] block diagram of ICA algorithm proposed in (69), where the whole dose delivery history and anatomy model from daily CBCT images are considered. This is a special case of **(A)** by taking the system state $\psi^{(t)} = (\psi^{(t)}_{cumdose}, \psi^{(t)}_{geometry})$.

a beamlet, $\Delta_{v,b,x}$ is a matrix computed for the snapshots of the anatomy in each phase, and θ_v, γ are some constants specific to the problem in question. The main purpose is to learn an optimal probability $p(x)$ as a stochastic control overall breathing phases $x \in X$ via Equation (29). The motion p.d.f. searched in the infinite-dimensional controls was actually approximated by the discretized set,

$$P = \{p \in \mathcal{F}(X;\mathbb{R}) \cong \mathbb{R}^{|X|} \,|\, p(x) \geq 0, \sum_{x\in X} p(x) = 1\} \tag{30}$$

such that this problem is tractable. They further required the **realization** $\tilde{p}$ of p in Equation (29) during a treatment to be constrained within certain error bounds ℓ and u:

$$\begin{aligned} P_U := \{\tilde{p}(x) \in P \,|\, & \underbrace{p(x) - \underline{p}(x)}_{\ell(x)} \leq \tilde{p}(x) \\ & \leq \underbrace{\underline{p}(x) + \tilde{p}(x)}_{u(x)}, \quad \forall x \in U \subseteq X\} \end{aligned} \tag{31}$$

As a result, the experiments by Bortfeld et al. showed that even when they allowed an unaccepted underdosage in the tumor anywhere between 6 and 11%, their proposal Equation (29) still offered same level of protection as the margin solution within 1% under dosage on average. Their approach proves that using stochastic controls helps stabilize the system with uncertainty over time. Later in (71), Chan and Mišić further improved the previous adaptive approach by extending the static probability distribution $\{p\}$ into a temporal sequence of PDF $(p^{(1)}, p^{(2)}, \ldots, p^{(k)})$ by incorporating uncertainty set updated each time for ART, which corresponds to the sequential control $\{\mathbf{u}^{(t)} | \, t \in T\}$ in Equation (18). The proposal in (71) essentially replaces the uncertainty p.d.f. $p \in P_U$ of Equation (29) by $p^{(k)} \in P_U^{(k)}$ iteratively to take care of patient's breathing motions.

$$\begin{aligned} & p^{(k+1)} \leftarrow p^{(k)}, \quad \text{with} \\ & p^{(k)} \in P_U^{(k)} := \{\tilde{p}(x) \in P \,|\, \ell^{(k)}(x) \leq \tilde{p}(x) \leq u^{(k)}(x), \quad \forall x \in U \subseteq X\} \end{aligned} \tag{32}$$

Two versions of uncertainty updates are proposed,

$$\ell^{(k+1)} = (1-\alpha)\,\ell^{(k)} + \alpha\, p^{(k)}, \quad u^{(k+1)} = (1-\alpha)\,u^{(k)} + \alpha\, p^{(k)} \tag{33}$$

$$\begin{aligned} \ell^{(k+1)} &= \frac{1}{k+1}\left(\ell^{(k)} + \sum_{i=1}^{k} p^{(i)}\right), \\ u^{(k+1)} &= \frac{1}{k+1}\left(u^{(k)} + \sum_{i=1}^{k} p^{(i)}\right) \end{aligned} \tag{34}$$

where the first version is called the *exponential* smoothing update and the second is called the *running average* update. Together, Equations (29) (32), (33), or (34) constituted their proposal in (71) and suggested that their method does not require accurate information to exist before a treatment commences. Their evaluation further stressed its clinical value as it allows for the tumor dose to be safely escalated without leading to additional healthy tissue toxicity, which may ultimately improve the rate of patient survival. Subsequently, Mar and Chan (72) further proposed an extension to the adaptive robust ART mentioned above (70, 71) by adding drift component using the Lujan model (73) of patients' breathing patterns.

Another related approach utilizing the formulation Equation (18) is found in (74), where Löf et al. developed statistical models for ART. Their design used *stochastic optimization* to handle two kinds of errors: (1) errors due to internal motion and change of organs (or tissues) and (2) errors due to the uncertainty in the geometrical setup of a patient. They attempted to compensate for the systematic errors by couch corrections and for the random error by modulation of the fluence profiles. This system was further modified by Rehbinder et al. using a linear–quadratic regulator (LQR) (75).

4.4. Reinforcement Learning (RL) for ART

RL is a set of machine learning algorithms that can interact with an "environment" (e.g., radiotherapy). Usually, there is a goal set for the RL, acting as an agent, to reach. Examples could be, winning a chess/board game or driving safely through a trip in an autonomous driving vehicle. Such a procedure is usually done by collecting the so-called *reward* designed by humans. RL serves as an independent machine learning area besides the common supervised or unsupervised learning mentioned earlier. RL is based on the environment defined by a Markov decision process (MDP).

An MPD is a 5-tuple $(S, \mathcal{A}, P, \gamma, R)$, where

- $S = \{(x_1, \ldots, x_n) \in \mathbb{R}^n\}$ is the space of all possible states,
- $\mathcal{A}$ is a finite collection of all (discrete) actions,
- $R : \Omega \to \mathbb{R}$ is the reward function given on the product space $\Omega = S \times \mathcal{A} \times S$,
- $\gamma \in [0, 1]$ is the discount factor, representing the importance (rewards) that propagates from the future back to the present,
- $P : \mathcal{F} \to [0, 1]$ is a probability measure on Ω with $\mathcal{F} = 2^{\Omega}$ the power set (σ-algebra) of Ω, whose probability mass function (pmf) $(s, a, t) \mapsto P(s, a, t)$ denotes the transition probability from state $s \in S$ to another $t \in S$ under an action $a \in \mathcal{A}$. Consequently, this induces the condition probability

$$P_{sa}(t) \equiv \text{Prob}(t\,|s, a) \equiv P(s, a, t)/P(s, a), \tag{35}$$

on space of next states t conditioned on previous state s and current action a.

As an example, in chess, each $s_i \in S$ will stand for a configuration of the chess board and action $a_i \in \mathcal{A}$ corresponds to a move given by a player. The purpose of an agent in the RL is to find a sequence of actions $\{a_0, a_1, \ldots\}$ (acting on an initial state $s_0 \in S$) such that a path in S collects maximum rewards (and hence winning the goal/game):

$$s_0 \xrightarrow[\pi]{a_0} s_1 \xrightarrow[\pi]{a_1} s_2 \xrightarrow[\pi]{a_2} s_3 \ldots \tag{36}$$

An agent is, by itself, a policy function π: $S \to \mathcal{A}$ who determines an action $a = \pi(s)$ under a state s, as described in Equation

(36). There are mainly two ways to construct a policy function by *policy-based* and *value-based* methods in RL: the former parametrizes a policy function directly (76) via $\pi^{(\theta)}$ while the latter builds one implicitly via Q-functions, and hence is also called Q-learning. The policy-based method is usually applied in continuous controls where $\mathcal{A} \cong \mathbb{R}$ or large cardinality $|\mathcal{A}| \to \infty$. In this study, we shall focus more on the Q-learning and its application in radiotherapy.

An optimal policy $\pi^*: S \to \mathcal{A}$ is derived from maximizing the Q-function in the Q-learning, such that $Q^{\pi^*} = \max_\pi Q^\pi$, where the Q-function is defined by evaluating the value at $(s, a) \in S \times \mathcal{A}$ via rewards collected in all possible paths:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R\left(s_k, \pi\left(s_k\right)\right) \middle| \pi,\ s_0 = s,\ a_0 = a\right] \tag{37}$$

However, this definition Equation (37) is ideal for comprehension, yet, difficult for actual computation. A practical realization of computing the Q-function is via the following Bellman's iteration, whose optimal value Q^{π^*} is computed by an iterative (functional) sequence $\{\tilde{Q}_i\}_{i=1}^{\infty}$ instead,

$$\tilde{Q}_{i+1}(s,a) = \mathbb{E}_{t\sim P_{sa}}\left[R(s,a) + \gamma \max_{b\in A} \tilde{Q}_i(t,b)\right]. \tag{38}$$

Such an iteration Equation (38) is guaranteed to converge by the contraction mapping theorem (77) of the uniquely fixed point as $\{\tilde{Q}_i\}_{i=1}^{\infty} \to Q^{\pi^*}$ if $i \to \infty$ (78) such that

$$\tilde{Q}^*(s,a) = \mathbb{E}_{t\sim P_{sa}}\left[R(s,a) + \gamma \max_{b\in A} \tilde{Q}^*(t,b)\right]. \tag{39}$$

The calculation soon becomes intractable when either the cardinality $|S|$ or $|\mathcal{A}|$ is large. A possible solution to this is utilizing deep learning methods for evaluating the Q-function proposed by Google DeepMind (79, 80), hence the name Deep Q-network (DQN). By taking advantages of neural networks, the convergence of the Q-function with Equation (38) becomes more efficient and accurate. DQN proposes $\tilde{Q}_i = Q_{\text{DNN}}^{\Theta_i}$, where Θ_i denotes the parametrization (weights) of the DNN at ith iteration and requires the following loss function being optimized:

$$\mathcal{L}_i(\Theta_i) = \mathbb{E}_{(s,a)\sim\rho} \times \left[\left(\mathbb{E}_{t\sim P_{sa}}\left[R(t,a) + \gamma \max_{b\in A} Q_{\text{DNN}}^{\Theta_{i-1}}(t,b)\right] - Q_{\text{DNN}}^{\Theta_i}(s,a)\right)^2\right]. \tag{40}$$

In short, Equations (38) and (40), and $\tilde{Q}_i = Q_{\text{DNN}}^{\Theta_i}$ together makes the DQN.

4.5. Example: Adapting RT Plans Using Deep Reinforcement Learning

Using the NSCLC dataset from Section 2.5.1, we attempt to apply a DQN to provide automatic dose escalation at the 2/3 period (about 4 weeks) into a treatment as illustrated in **Figure 13**, where the dose escalation is the action to be submitted by the DQN. The main goal of the study is to compare the automatic decision made by the DQN to that established by a clinical protocol (81). This will be described briefly in the following, details can be consulted in (23).

That work explicitly presented a suitable MDP. In particular, a state space chosen to be useful for prediction of local control (LC) and RP2 based on the BN formalism introduced in Section 3.3.2.

By defining the state space as $S = \{(x_1, \ldots, x_n) \in \mathbb{R}^n\}$ with $n = 9$ and

$$\begin{aligned} &x_1 = \text{IL4} \quad x_2 = \text{IL15}, \quad x_3 = \text{GLSZM.GLN}, \ x_4 = \text{GLRLM.RLN}, \\ &x_5 = \text{MCP1}, \ x_6 = \text{TGF}\beta 1, \ x_7 = \text{Lung gEUD}, \quad x_8 = \text{Tumor gEUD}, \ x_9 = \text{MTV} \end{aligned} \tag{41}$$

where x_1, x_2, x_5, x_6, x_9 are cytokines, x_3, x_4 are of PET radiomics, and x_7, x_8 are doses, and here, the allowed action set will be $\mathcal{A} = \{a_1 = \text{dose/frac}\} \subseteq \mathbb{R}^+$. One notices that such a choice of a MDP for dose automation is not unique; there may exist other environments to attain the same or even better performance (82).

A tricky problem is that the transition probability in Equation (35) is intractable to the real world (radiotherapy environment); therefore, DNNs were utilizes to model the radiotherapy environment. Thus, a DNN provided an approximate transition probability $\tilde{P}(s,a;t) := \tilde{P}_{sa}(t) := \widetilde{\text{Prob}}(t \,|\, s,a)$ modeled from the observed data, where the transition takes place $s \xrightarrow{a} t$ under

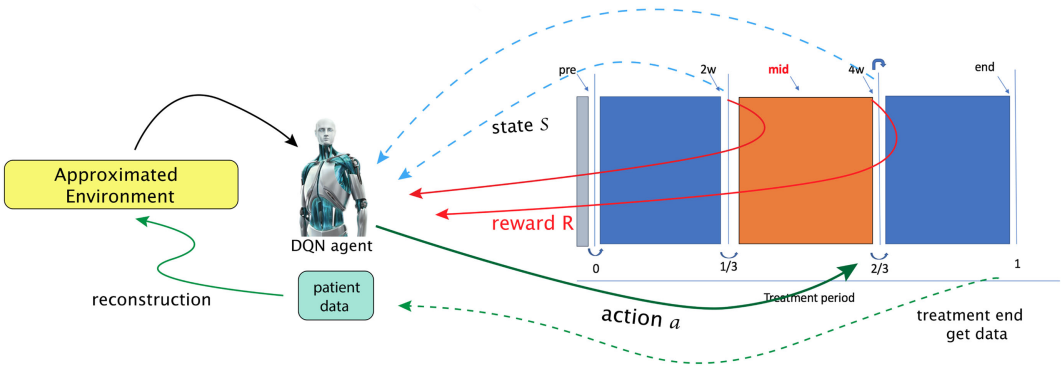

FIGURE 13 | In the paper (23), Tseng et al. proposed to utilize reinforcement learning for making decisions at 2/3 period of a treatment (right solid-green arrow). A first step in their framework is to learn transition functions from the historical data of two transitions recorded (RHS figure) so that the radiotherapy environment can be reconstructed (called *approximated environment*). With the transitions simulated, a DQN agent can then search for optimal dose at each stage [figures reprinted with permission].

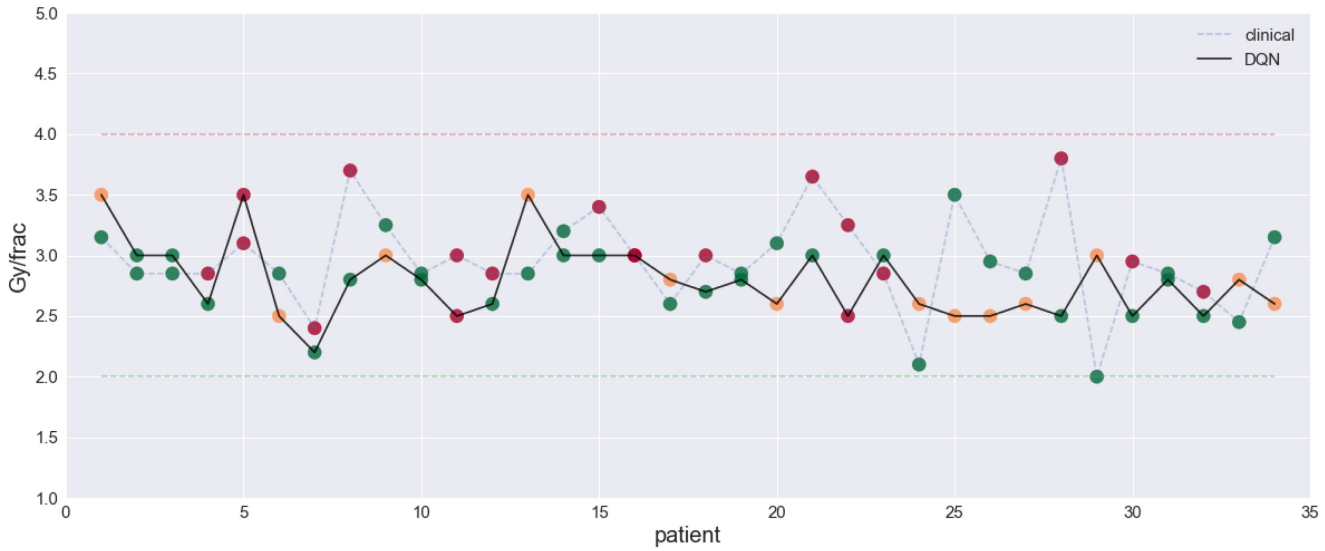

FIGURE 14 | This figure visualizes the dose fraction recommended by the clinicians (blue dashed line) and the autonomous DQN (black solid line). Differences and similarities can thus be compared, with RMSE = 0.5 Gy. An evaluation of good (green dots), bad (red dots), and *potentially good* decisions (orange dots) (23) [figures reprinted with permission].

action a. Another problem to solve in that the sample size was small relative to the DNNs. Hence, a Generated Adversarial Network (GAN) technique was used to alleviate this problem.

After proper choice of actions, $\mathcal{A} = \{1, 1.1, 1.2, \ldots, 5\}$ Gy, and a reward function looking upon to higher LC than usual P^+ baseline function:

$$R(s) = \frac{1}{2}\sqrt{\text{Prob}(\text{LC}\,|\,s)} \cdot (1 - 0.8 \cdot \text{Prob}(\text{RP2}\,|\,s)) \cdot (1 + sgn(17.2\% - \text{Prob}(\text{RP2}\,|\,s))), \quad (42)$$

The results demonstrated the feasibility to derive automated dose levels (black solid line) that are similar to or compatible with the clinical protocol (blue dashed line) as shown in **Figure 14** with the corresponding statistics shown in **Table 1**.

5. DISCUSSION

5.1. Statistical and Probabilistic Aspects

Here, we attempt to provide a fundamental statistical and probabilistic interpretation for sequential machine learning algorithms to help understand their roles in KBR-ART. This will be done with the specific focus on how knowledge can accumulates in such a KBR-ART system when the known information in the system is growing with time. First, we characterize the probability space as: $(\Omega, \mathcal{F}, \mathcal{P})$, where $\mathcal{F}$ is a σ-algebra[1] of a sample space Ω and $P : \mathcal{F} \to \mathbb{R}^+$ is the probability measure defined on Ω, see (83, 84). In this setting, Ω denotes the set of all possible outcomes and $\mathcal{F}$ as the space of all events. A (multi-dimensional) random variable $\mathbf{X}$ is a then $\mathcal{F}$-measurable function $\mathbf{X} : \Omega \to \mathbb{R}^n$ on a probability space $(\Omega, \mathcal{F}, P)$. Roughly speaking, the σ-algebra corresponds to the "information" useful (and related) to the random variable $\mathbf{X}$. Furthermore, if $\{\mathbf{X}^{(t)}|\, t \in T\}$ is a sequence of random variables (or a process), a natural σ-algebra induced by the process is defined by:

$$\mathcal{U}(t) := \mathcal{U}(\mathbf{X}(s)|\, s \in [0, t]) := \left\{ \left(\mathbf{X}^{(s)}\right)^{-1}(B) \subseteq \Omega |\, \forall \text{ Borel set } B \subseteq \mathbb{R}^n,\ \forall s \in [0, t] \right\}, \quad (43)$$

which is interpreted as the *history* of the process up to time t. Therefore, under a process $\{\mathbf{X}^{(t)}|\ t \in T\}$, one can regard the σ-algebra $\mathcal{U}(t)$ as accumulating information from the observed variable $\mathbf{X}^{(s)}$ along the times $s \in [0, t]$. Thus, a one liner may be best to represent the message we try to deliver:

a σ-algebra = information;

a "growing" σ-algebra = more information coming in.

In fact, the idea of considering growing information, such as weather forecasting, stock pricing prediction, or daily CT changes, can be understood by a growing σ-algebra called a *filtration*, **Figure 15**. Such tool for analysis is commonly seen in quantitative finance (85, 86), which we believe it shares the same nature as a treatment in radiotherapy. The following concept describes growing (accumulating) information.

A sequence of σ-algebras $\{\mathcal{F}_t\}_{t \geq 0}$ on a measurable space $(\Omega, \mathcal{F})$ with $\mathcal{F}_t \subseteq \mathcal{F}$ is called a filtration if $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$ whenever $t_1 \leq t_2$.

TABLE 1 | Summary for the evaluation on clinicians' and the DQN decisions extracted from (23).

Summary	Good	Bad	Potentially good
Clinicians	19 (55.9%)	15 (44.1%)	0
DQN	17 (50%)	4 (11.8%)	13 (38.2%)

[1] $\mathcal{F}$ as a collection of subsets of a set Ω is called a σ-algebra if the following three is satisfied: (1) $\Omega \in \mathcal{F}$, (2) if $A \in \mathcal{U}$ implies $(\Omega \setminus A) \in \mathcal{F}$, and (3) arbitrary union $A = \cup_{k=1}^{\infty} A_k \in \mathcal{F}$ if $A_k \in \mathcal{F}$.

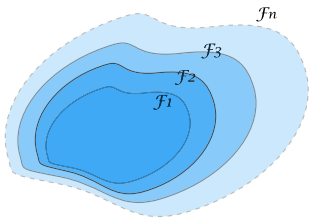

FIGURE 15 | An illustration of a filtration indicating a sequence of growing σ-algebras, $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$ if $t_1 \leq t_2$. The enlargement of σ-algebras reflects the accumulating information as time evolves. This provides a theoretical tool to measure the growth of knowledge in a KBR-ART design.

The labeling index t is usually referred to "time" or a similar concept, where in the radiotherapy case it may be treatment fractions, stages, or phases. If we consider a filtration generated from a stochastic process via $\mathcal{F}_t = \mathcal{U}(t)$, then, intuitively, this filtration is interpreted as containing all history available up to time t, but not future information available about the process. Due to this nature, a process adapted to a filtration $\mathcal{F}$ is also called *non-anticipating*, indicating that one cannot see into the future.

Therefore, a KBR-ART system would rely on machine learning algorithms (such as CNN, RNN, DRL, ... etc.) to explore non-anticipating filtrations and to learn from accumulating knowledge or information, such as the examples givens in Sections 3.3 and 4.

To demonstrate the concept of filtrations more concretely in our setting, the following example is provided. Suppose a sequence of independent random variables $\{X^{(i)}\}_{i=1,2,3...}$ denotes the *growth* in GTV size at stage i with $\mathbb{E}(X^{(i)}) = d_i$ for all i. If we have measured total growth up to stage k, i.e., $S^{(k)} := X^{(1)} + \ldots + X^{(k)}$, we like to know what is our best guess for the growth after n more stages $S^{(k+n)}$, given the information of the past $S^{(1)}, \ldots, S^{(n)}$?

Some computation reveals that

$$\mathbb{E}(S^{(k+n)} | S^{(1)}, \ldots, S^{(k)}) = \mathbb{E}\left(X^{(1)} + \cdots + X^{(k+n)} | S^{(1)}, \ldots, S^{(k)}\right) = S^{(k)} + \sum_{i=k+1}^{n} d_i, \tag{44}$$

which indicates that the best surmise for the future value $S^{(k+n)}$, given the knowledge (history) up to stage k, is $S^{(k)}$ plus empirical understanding (averages), reflecting the information cease to grow after time step k. The computation Equation (44) relies on the following fact:

1. *If X is $\mathcal{F}$-measurable, then $\mathbb{E}(X|\mathcal{F}) = X$ almost surely.*
2. *If X is independent of $\mathcal{F}$, then $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(x)$ almost surely*

After the above discussion of how information can be accumulated using σ-algebras, next, we discuss how to analyze sequential random variables from a more theoretical perspectives using time series.

5.1.1. Time Series

Due to the nature of sequential data, an KBR-ART is naturally related to time series, which are applied comprehensively in forecasting, such as econometrics, quantitative finance, seismology, and signal processing, etc. Quoting from (87):

> *A time series model for the observed data $\{\boldsymbol{x}^{(t)} | t \in T\}$ is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X^{(t)} | t \in T\}$ of which $\{\boldsymbol{x}^{(t)}\}$ is postulated to be a realization.*

Incidentally, a time series is a special case of **stochastic processes** $\{\mathbf{X}^{(t)} | t \in T\}$, where the time labeling set T can be an infinite set. In a very general case, a process $\{X^{(t)} | t \in \mathbb{Z}\}$ can have *Volterra expansion*

$$X^{(t)} = c + \sum_{j=0}^{\infty} \vartheta_j Z^{(t-j)} + \sum_{j,k}^{\infty} \vartheta_{jk} Z^{(t-j)} Z^{(t-k)} + \sum_{j,k,\ell}^{\infty} \vartheta_{jk} Z^{(t-j)} Z^{(t-k)} Z^{(t-\ell)} + \cdots, \tag{45}$$

where high order terms can be considered. Usually, the modeling of time series is divided by two main categories, linear and non-linear methods.

In particular, there are three classes of linear models that carry practical importance, namely autoregressive models AR(p), the moving average models MA(q), and the integrated (I) models.

(The ARMA (p,q) process with mean μ) The process $\{X^{(t)} | t \in \mathbb{Z}\}$ is called an ARMA (p,q) process if it is stationary and satisfies for all t,

$$\left(X^{(t)} - \mu\right) - \varphi_1 \left(X^{(t-1)} - \mu\right) - \cdots - \varphi_p \left(X^{(t-p)} - \mu\right) = Z^{(t)} - \vartheta_1 Z^{(t-1)} - \cdots - \vartheta_q Z^{(t-q)}, \tag{46}$$

where $\mu, \varphi_i, \vartheta_i \in \mathbb{R}$ and $\{Z^{(t)}\} \simeq \mathrm{WN}\left(0, \sigma^2\right)$ are white noise (error terms).

Here, the ARMA(p, q) process refers to the model with p *autoregressive* terms and q *moving-average* terms. Especially, $p = 0$ and $q = 0$ in the ARMA(p, q) process corresponds to two useful linear cases called AR(p) and MA(q) models, respectively. The aim of studying the behavior of a time series $\{X^{(t)}\}$ can be done *via* the analysis of the depending coefficients φ_I, ϑ_i and its autocorrelation function (88), which we will not go through. An interesting fact is that one can study the causality of an ARMA(p, q) process via the following fact:

Let $\{X^{(t)}\}$ be an ARMA(p, q) process with $\varphi(z) := (1 + \varphi_1 z + \cdots + \varphi_1 z^p)$, $\vartheta(z) := (1 + \vartheta_1 z + \cdots + \vartheta_q z^q)$ have no common zeros. Then $\{X^{(t)}\}$ is causal if and only if $\varphi|_{\mathcal{D}} \neq 0$ with $\mathcal{D} = \{z \in \mathbb{C} | \parallel z \parallel \leq 1\}$.

Thus AR(1) process with $\mu = 0$ is only a simple case given by $X^{(t)} = Z^{(t)} + X^{(t-1)}$ from Equation (46). Since $\varphi(z) = 1 - \varphi_1 z$, it follows that $\{X^{(t)}\}$ is causal if $|\varphi_1| < 1$ and non-stationary when $|\phi_1| = 1$. This AR(1) case demonstrates that we may actually learn the behavior of a time series by analyzing the dependent coefficients. In fact, the heuristic AR(1) process is directly related to the Markov process due to a fact (see Proposition 7.6 in (89)). Simply stated, for a process $\{X^{(t)} | t \in \mathbb{Z}\}$ taking values in a Borel space S, Z_1, Z_2,... are independent taking values in E and if there exist functions f_t: $S \times E \to S$, $t \in \mathbb{Z}$, such that $X^{(t)}$ is recursively defined by

$$X^{(t)} = f_t\left(X^{(t-1)}, Z^{(t)}\right), \qquad X^{(0)} = x_0 \in S, \tag{47}$$

then the process $\{X^{(t)} | t \in \mathbb{Z}\}$ is Markov. This result Equation (47) then justifies the claim that the AR(1) is a Markov process as the transition functions simply indicate $f_t(X^{(t-1)}, Z^{(t)}) = Z^{(t)} + \varphi_1 X^{(t-1)}$ from Equation (47). Moreover, it is time-homogeneous since $\{Z^{(t)}\}$ are i.i.d. and f_t is fixed across all t. As one recalls that the Markov process is defined under the property

$$P\left(\left(X^{(t)}\right)^{-1}(B) \,|\, \mathcal{U}(s)\right) = P\left(\left(X^{(t)}\right)^{-1}(B) \,|\, X^{(s)}\right) \quad (\forall \text{ Borel } B \subseteq \mathbb{R}^n,\ t \geq s \geq 0) \tag{48}$$

where $\mathcal{U}(s)$ is as defined in Equation (43). At the prediction level, AR(1) or Markov process then indicates that one can estimate the probabilities of future values $X^{(t)}$ just as well as if one was aware of the entire history of the process $\mathcal{U}(s)$ prior to time s. The Markov property Equation (48) serves as a simplifying assumption to reduce complexities in variables involved. Therefore, it is one of our reasons to introduce the Bayesian Networks modeling based on Markov process in Section 3.2.4.

5.2. Comparison of Varying Data-Driven Models

There are a large number of statistical models in the area of machine learning. They can be basically divided into 3 categories: *supervised, unsupervised*, and *reinforcement learning*, where supervised models are mainly used for data *prediction*, unsupervised models are usually used to explore intrinsic data structure such as probability and location distribution, and the reinforcement learning, which we will introduce in Section 4.4 is to learn best controls within certain circumstances. All the methods introduced in Section 3.2 belong to the supervised learning category, which is the cornerstone for KBR-ART system implementation. It is essential for a KBR-ART to have a accurate model for future prediction in patients' status, e.g., organ geometry and shape changing, whether the model is analytical or statistical. Statistical modeling is typically a handy choice over analytical one to overcome the modeling complexity involved in mechanistic realizations of radiotherapy interactions.

Comparison of the merits of several classical methods such as linear regression Section 3.2.1, Bayesian networks Section 3.2.4, decision trees, and SVMs can be found in (90–92). Generally speaking, the pros of classical data-driven models such as linear regression and Bayesian networks is that they are interpretable, numerically stable, computational efficient, and work even on small sample-sized dataset, but the cons are that they lack versatility in tasking (e.g., no one uses regressions for image segmentation or contouring) and do not possess the ability to handle complex and high variety of data, such as images, video, sequences, languages, and mixture data. For complex data such as the RT data, one can rely on more modern techniques such as deep learning, particularly DNN, CNN, and RNN-based structures. For intensive review regarding deep learning and their merits, one may refer to (42, 93). The trade-off between handling complex data and data interpretability may drive one to choose between classical and deep machine learning methods. Moreover, deep learning techniques typically require larger amount of observations compared to classical statistical learning techniques. This is a main reason that deep learning is no yet as prominent in medical and biological field compared to its current dominant in computer science and engineering. The bottom line here is that there yet no universal recognition for which classifier can do the best job in biomedicine or oncology. The development of KBR-ART is foreseeable to rely more deep learning approaches for outcome modeling and variety tasks of (image, sequential) data processing and decision-making.

6. CONCLUSION

In this study, we presented a framework for comprehensive KBR-ART design and implementation based on machine learning and explored some of its main characteristics. First, in Section 2, we analyzed the characteristics and types of features in clinical data as effective choice of data for feeding knowledge into KBR-ART. Second, in Section 3, we visited a few promising and powerful techniques of modern machine learning development, such as DNNs, CNNs, RNNs as well as the classical linear regression-type models. The KBR-ART framework we proposed here rely on machine learning techniques, which are capable of accurate prediction and sequential learning, which are the cornerstones for building up a KBR-ART system. There are three pertained questions to the design and realization of KBR-ART, which we addressed in this paper and we presented illustrative examples for each case highlighted by the application RL/BN onto a NSCLC radiotherapy dataset. In Section 4, we provided a unifying formulation in Section 4.1 for designing a KBR-ART system (Equation 18). The purpose was twofold: (1) to clearly understand the essence of previous constructed ARTs of last generation, (2) to provide a guiding principle for designing next generation algorithms.

The application of the presented technologies here provides great promise for the field of KBR-ART, yet there are still numerous challenges ahead. First, there is highly complex nature of radiation interaction with human biology that we are still trying to develop a better understanding. Second, medical datasets typically suffer from small sizes and often incomplete. Several efforts between nations and domestic institutes are being carried out to consolidate larger datasets for oncology studies, for the purpose of statistical model training and validations, but many are still in the infancy. Nevertheless, this paper still serves as a blueprint laying the foundation for the establishment and applicability of KBR-ART using modern machine learning techniques.

AUTHOR CONTRIBUTIONS

H-HT writes up and collects main materials related to the study; YL also collects materials related to the study; RTH plots and organizes the study; IEN directs and organizes the study.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jolly and Dr. Kong for the help of providing the lung testing datasets for the study. H-HT thanks Sunan Cui for numerous fruitful discussions. This work was supported in part by the National Institutes of Health P01 CA059827.

REFERENCES

1. Stanley HB, El Naqa I, Klein EE. Introduction to big data in radiation oncology: exploring opportunities for research, quality assessment, and clinical care. *Int J Radiat Oncol Biol Phys* (2016) 95(3):871–2. doi:10.1016/j.ijrobp.2015.12.358
2. El Naqa I. Perspectives on making big data analytics work for oncology. *Methods* (2016) 111:32–44. doi:10.1016/j.ymeth.2016.08.010
3. Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online adaptive radiation therapy. *Int J Radiat Oncol Biol Phys* (2017) 99(4):994–1003. doi:10.1016/j.ijrobp.2017.04.023
4. Xing L, Siebers J, Keall P. Computational challenges for image-guided radiation therapy: framework and current research. *Semin Radiat Oncol* (2007) 17(4):245–57. doi:10.1016/j.semradonc.2007.07.004
5. Graves A, Schmidhuber J. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in Neural Information Processing Systems*. (2009). p. 545–52.
6. Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. (2014). p. 1764–72.
7. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* (1999) 15(11):937–46. doi:10.1093/bioinformatics/15.11.937
8. Übeyli ED. Recurrent neural networks with composite features for detection of electrocardiographic changes in partial epileptic patients. *Comput Biol Med* (2008) 38(3):401–10. doi:10.1016/j.compbiomed.2008.01.002
9. Übeyli ED. Combining recurrent neural networks with eigenvector methods for classification of ecg beats. *Digit Signal Process* (2009) 19(2):320–9. doi:10.1016/j.dsp.2008.09.002
10. Shen H, George D, Huerta E, Zhao Z. *Denoising gravitational waves using deep learning with recurrent denoising autoencoders*. (2017). *arXiv preprint arXiv:1711.09919*.
11. Wu JT, Dernoncourt F, Gehrmann S, Tyler PD, Moseley ET, Carlson ET, et al. Behind the scenes: a medical natural language processing project. *Int J Med Inform* (2018) 112:68–73. doi:10.1016/j.ijmedinf.2017.12.003
12. Marks LB. Dosimetric predictors of radiation-induced lung injury. *Int J Radiat Oncol Biol Phys* (2002) 54(2):313–6. doi:10.1016/S0360-3016(02)02928-0
13. Levegrün S, Jackson A, Zelefsky MJ, Skwarchuk MW, Venkatraman ES, Schlegel W, et al. Fitting tumor control probability models to biopsy outcome after three-dimensional conformal radiation therapy of prostate cancer: pitfalls in deducing radiobiologic parameters for tumors from clinical data. *Int J Radiat Oncol Biol Phys* (2001) 51(4):1064–80. doi:10.1016/S0360-3016(01)01731-X
14. Hope AJ, Lindsay PE, El Naqa I, Alaly JR, Vicic M, Bradley JD, et al. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *Int J Radiat Oncol Biol Phys* (2006) 65(1):112–24. doi:10.1016/j.ijrobp.2005.11.046
15. Bradley J, Deasy JO, Bentzen S, El Naqa I. Dosimetric correlates for acute esophagitis in patients treated with radiotherapy for lung carcinoma. *Int J Radiat Oncol Biol Phys* (2004) 58(4):1106–13. doi:10.1016/j.ijrobp.2003.09.080
16. Blanco AI, Chao KC, El Naqa I, Franklin GE, Zakarian K, Vicic M, et al. Dose-volume modeling of salivary function in patients with head-and-neck cancer receiving radiotherapy. *Int J Radiat Oncol Biol Phys* (2005) 62(4):1055–69. doi:10.1016/j.ijrobp.2004.12.076
17. Deasy JO, El Naqa I. *Image-Based Modeling of Normal Tissue Complication Probability for Radiation Therapy*. Boston, MA: Springer (2008). p. 211–52.
18. El Naqa I, Suneja G, Lindsay P, Hope AJ, Alaly J, Vicic M, et al. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. *Phys Med Biol* (2006) 51(22):5719. doi:10.1088/0031-9155/51/22/001
19. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* (2012) 48(4):441–6. doi:10.1016/j.ejca.2011.11.036
20. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. *Phys Med* (2017) 38:122–39. doi:10.1016/j.ejmp.2017.05.071
21. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* (2001) 69(3):89–95. doi:10.1067/mcp.2001.113989
22. El Naqa I, Craft J, Oh J, Deasy J. Biomarkers for early radiation response for adaptive radiation therapy. *Adapt Radiat Ther* (2011) 53–68.
23. Tseng H-H, Luo Y, Cui S, Chien J-T, Ten Haken RK, El Naqa I. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med Phys* (2017) 44(12):6690–705. doi:10.1002/mp.12625
24. Luo Y, El Naqa I, McShan DL, Ray D, Lohse I, Matuszak MM, et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother Oncol* (2017) 123(1):85–92. doi:10.1016/j.radonc.2017.02.004
25. Webb S. *The Physics of Three Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning*. CRC Press (1993). Available from: https://www.taylorfrancis.com/books/9781420050363
26. Joiner MC, Van der Kogel A. *Basic Clinical Radiobiology*. (Vol. 2). CRC Press (2016). Available from: https://www.crcpress.com/Basic-Clinical-Radiobiology-Fourth-Edition/Joiner-van-der-Kogel/p/book/9780340929667
27. El Naqa I. *A Guide to Outcome Modeling in Radiotherapy and Oncology: Listening to the Data*. CRC Press (2018).
28. Hall E, Giaccia A. *Radiobiology for the Radiologist*. Philadelphia: Lippincott Williams & Wilkins (2006).
29. Zaider M, Minerbo G. Tumour control probability: a formulation applicable to any temporal protocol of dose delivery. *Phys Med Biol* (2000) 45(2):279. doi:10.1088/0031-9155/45/2/303
30. Goitein M. Tumor control probability for an inhomogeneously irradiated target volume. *Eval Treat Plan Part Beam Radiother* (1987).
31. Lyman JT. Complication probability as assessed from dose-volume histograms. *Radiat Res Suppl* (1985) 8:S13–9. doi:10.2307/3576626
32. Niemierko A. Reporting and analyzing dose distributions: a concept of equivalent uniform dose. *Med Phys* (1997) 24(1):103–10. doi:10.1118/1.598063
33. Niemierko A. A generalized concept of equivalent uniform dose (eud). *Med Phys* (1999) 26(6):1100.
34. Coates J, Jeyaseelan AK, Ybarra N, David M, Faria S, Souhami L, et al. Contrasting analytical and data-driven frameworks for radiogenomic modeling of normal tissue toxicities in prostate cancer. *Radiother Oncol* (2015) 115(1):107–13. doi:10.1016/j.radonc.2015.03.005
35. Tucker SL, Li M, Xu T, Gomez D, Yuan X, Yu J, et al. Incorporating single-nucleotide polymorphisms into the lyman model to improve prediction of radiation pneumonitis. *Int J Radiat Oncol Biol Phys* (2013) 85(1):251–7. doi:10.1016/j.ijrobp.2012.02.021
36. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R*. (Vol. 112). New York: Springer (2013).
37. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol* (1996) 58(1):267–88.
38. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol* (2005) 67(2):301–20. doi:10.1111/j.1467-9868.2005.00527.x
39. Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the Twenty-First International Conference on Machine Learning*. New York, NY: ACM (2004). 78 p.
40. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw* (1989) 2(5):359–66. doi:10.1016/0893-6080(89)90020-8
41. Bovier A, Picco P. *Mathematical Aspects of Spin Glasses and Neural Networks*. (Vol. 41). Boston: Springer Science & Business Media (2012).
42. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. (Vol. 1). Cambridge: MIT press (2016).
43. Bulat I, Lei X. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* (2016) 44(2):547–57. doi:10.1002/mp.12045
44. Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol* (2017) 7:315. doi:10.3389/fonc.2017.00315
45. Qin W, Wu J, Han F, Yuan Y, Zhao W, Ibragimov B, et al. Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation. *Phys Med Biol* (2018) 63(9):095017. doi:10.1088/1361-6560/aabd19
46. Wang Y, Zu C, Hu G, Luo Y, Ma Z, He K, et al. Automatic tumor segmentation with deep convolutional neural networks for radiotherapy applications. *Neural Process Lett* (2018). doi:10.1007/s11063-017-9759-3

47. Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, et al. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Phys Med Biol* (2017) 62(21):8246. doi:10.1088/1361-6560/aa8d09
48. Luo Y, McShan D, Ray D, Matuszak M, Jolly S, Lawrence T, et al. Development of a fully cross-validated Bayesian network approach for local control prediction in lung cancer. *IEEE Transactions on Radiation and Plasma Medical Sciences*. (2018). p. 1–1. Available from: https://ieeexplore.ieee.org/abstract/document/8353476/
49. Ogunmolu OP, Gu X, Jiang SB, Gans NR. Nonlinear systems identification using deep dynamic neural networks. *CoRR* (2016) abs/1610.01439.
50. Cai J, Lu L, Xie Y, Xing F, Yang L. Pancreas segmentation in MRI using graph-based decision fusion on convolutional neural networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer International Publishing (2017). p. 674–82.
51. Roth HR, Lu L, Farag A, Sohn A, Summers RM. Spatial aggregation of holistically-nested networks for automated pancreas segmentation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Cham: Springer International Publishing (2016). p. 451–9.
52. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann Publishers Inc (1988).
53. Griffiths T, Yuille A. A primer on probabilistic inference. *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: Oxford University Press (2008) p. 33–57. doi:10.1093/acprof:oso/9780199216093.001.0001
54. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. New York: Springer (2012).
55. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn* (1997) 29(2–3):131–63. doi:10.1023/A:1007465528199
56. Kalet AM, Gennari JH, Ford EC, Phillips MH. Bayesian network models for error detection in radiotherapy plans. *Phys Med Biol* (2015) 60(7):2735. doi:10.1088/0031-9155/60/7/2735
57. Gomes E, Duarte J, Frutuoso e Melo PF. Human reliability modeling of radiotherapy procedures by bayesian networks and expert opinion elicitation. *Nucl Technol* (2016) 194(1):73–96. doi:10.13182/NT15-29
58. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. *Med Phys* (2010) 37(4):1401–7. doi:10.1118/1.3352709
59. Oh JH, Craft J, Lozi RA, Vaidya M, Meng Y, Deasy JO, et al. A bayesian network approach for modeling local failure in lung cancer. *Phys Med Biol* (2011) 56(6):1635. doi:10.1088/0031-9155/56/6/008
60. Lee S, Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P, et al. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. *Med Phys* (2015) 42(5):2421–30. doi:10.1118/1.4915284
61. Luo Y, El Naqa I, McShan D, Matuszak M, Jolly S, Haken RKT. Simultaneous prediction of specific radiotherapy outcomes using a multi-objective bayesian network (moBN) approach. *Int J Radiat Oncol Biol Phys* (2017) 99(2, Suppl):S35. doi:10.1016/j.ijrobp.2017.06.094
62. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and validating a survival prediction model for nsclc patients through distributed learning across 3 countries. *Int J Radiat Oncol Biol Phys* (2017) 99(2):344–52. doi:10.1016/j.ijrobp.2017.04.021
63. Tucker SL, Zhang M, Dong L, Mohan R, Kuban D, Thames HD. Cluster model analysis of late rectal bleeding after imrt of prostate cancer: a case-control study. *Int J Radiat Oncol Biol Phys* (2006) 64(4):1255–64. doi:10.1016/j.ijrobp.2005.10.029
64. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *CoRR* (2014).
65. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision (ICCV)*. (2017). p. 618–26. Available from: https://ieeexplore.ieee.org/document/8237336/
66. Ågren A, Brahme A, Turesson I. Optimization of uncomplicated control for head and neck tumors. *Int J Radiat Oncol Biol Phys* (1990) 19(4):1077–85. doi:10.1016/0360-3016(90)90037-K
67. Keller H, Ritter MA, Mackie T. Optimal stochastic correction strategies for rigid-body target motionoptimal stochastic correction strategies for rigid-body target motion. *Int J Radiat Oncol Biol Phys* (2003) 55(1):261–70. doi:10.1016/S0360-3016(02)03867-1
68. Humpherys J, Redd P, West J. A fresh look at the kalman filter. *SIAM Rev* (2012) 54(4):801–23. doi:10.1137/100799666
69. de la Zerda A, Armbruster B, Xing L. Formulating adaptive radiation therapy (ART) treatment planning into a closed-loop control framework. *Phys Med Biol* (2007) 52(14):4137. doi:10.1088/0031-9155/52/14/008
70. Bortfeld T, Chan TCY, Trofimov A, Tsitsiklis JN. Robust management of motion uncertainty in intensity-modulated radiation therapy. *Oper Res* (2008) 56(6):1461–73. doi:10.1287/opre.1070.0484
71. Chan TC, Mišić VV. Adaptive and robust radiation therapy optimization for lung cancer. *Eur J Oper Res* (2013) 231(3):745–56. doi:10.1016/j.ejor.2013.06.003
72. Mar PA, Chan TCY. Adaptive and robust radiation therapy in the presence of drift. *Phys Med Biol* (2015) 60(9):3599. doi:10.1088/0031-9155/60/9/3599
73. Lujan AE, Larsen EW, Balter JM, Ten Haken RK. A method for incorporating organ motion due to breathing into 3d dose calculations. *Med Phys* (1999) 26(5):715–20. doi:10.1118/1.598577
74. Löf J, Lind BK, Brahme A. An adaptive control algorithm for optimization of intensity modulated radiotherapy considering uncertainties in beam profiles, patient set-up and internal organ motion. *Phys Med Biol* (1998) 43(6):1605. doi:10.1088/0031-9155/43/6/018
75. Rehbinder H, Forsgren C, Löf J. Adaptive radiation therapy for compensation of errors in patient setup and treatment delivery. *Med Phys* (2004) 31(12):3363–71. doi:10.1118/1.1809768
76. Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: Solla SA, Leen TK, Müller K, editors. *Advances in Neural Information Processing Systems*. MIT Press (2000). p. 1057–63. Available from: http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation.pdf
77. Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. MIT Press (2012).
78. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. (Vol. 1). Cambridge: MIT press (1998).
79. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. *Playing atari with deep reinforcement learning*. (2013). *arXiv preprint arXiv:1312.5602*.
80. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature* (2015) 518(7540):529–33. doi:10.1038/nature14236
81. Kong FM, Ten Haken RK, Schipper M, Frey KA, Hayman J, Gross M, et al. Effect of midtreatment PET/CT-adapted radiation therapy with concurrent chemotherapy in patients with locally advanced non–small-cell lung cancer: a phase 2 clinical trial. *JAMA Oncol* (2017) 3(10):1358–65. doi:10.1001/jamaoncol.2017.0982
82. Schuck NW, Wilson RC, Niv Y. *A State Representation for Reinforcement Learning and Decision-Making in the Orbitofrontal Cortex*. (2017). *bioRxiv*.
83. Chung KL, editor. Probability and mathematical statistics. *A Course in Probability Theory (Second Edition)*. San Diego: Academic Press (1974). Available from: http://www.sciencedirect.com/science/article/pii/B978080057040250001X
84. Durrett R. *Probability: Theory and Examples*. Cambridge University Press (2010). Available from: http://www.cambridge.org/us/academic/subjects/statistics-probability/probability-theory-and-stochastic-processes/probability-theory-and-examples-4th-edition?format=HB&isbn=9780521765398
85. Davis M. Mathematics of financial markets. In: Engquist B, Schmid W, editors. *Mathematics Unlimited—2001 and Beyond*. Berlin, Heidelberg: Springer (2001). p. 361–80.
86. Privault N. *Stochastic Analysis in Discrete and Continuous Settings: With Normal Martingales*. Berlin, Heidelberg: Springer (2009).
87. Brockwell PJ, Davis RA. *Introduction to Time Series and Forecasting*. Switzerland: Springer International Publishing (2016).
88. Brockwell PJ, Davis RA, Fienberg SE, Berger JO, Gani J, Krickeberg K, et al. *Time Series: Theory and Methods*. New York: Springer (1991).
89. Kallenberg O. *Foundations of Modern Probability*. New York: Springer (2006).
90. Kotsiantis SB. Supervised machine learning: a review of classification techniques. *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications*

in eHealth, HCI, Information Retrieval and Pervasive Technologies. Amsterdam, Netherlands: IOS Press (2007). p. 3–24.
91. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. *Knowl Inform Syst* (2008) 14(1):1–37. doi:10.1007/s10115-007-0114-2
92. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*. New York, NY: ACM (2006). p. 161–8.
93. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* (2015) 521(7553):436. doi:10.1038/nature14539

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Oncology

ORIGINAL RESEARCH
published: 17 August 2018
doi: 10.3389/fonc.2018.00294

Machine Learning Applications in Head and Neck Radiation Oncology: Lessons From Open-Source Radiomics Challenges

Hesham Elhalawani[1†], Timothy A. Lin[1,2†], Stefania Volpe[1,3], Abdallah S. R. Mohamed[1,4], Aubrey L. White[1,5], James Zafereo[1,5], Andrew J. Wong[1,6], Joel E. Berends[1,6], Shady AboHashem[1,7], Bowman Williams[1,8], Jeremy M. Aymard[1,9], Aasheesh Kanwar[1,10], Subha Perni[1,11], Crosby D. Rock[1,12], Luke Cooksey[1,13], Shauna Campbell[1,14], Pei Yang[1,2], Khahn Nguyen[15], Rachel B. Ger[16,17], Carlos E. Cardenas[16,17], Xenia J. Fave[18], Carlo Sansone[19], Gabriele Piantadosi[19], Stefano Marrone[19], Rongjie Liu[2,20], Chao Huang[2,20], Kaixian Yu[2,20], Tengfei Li[2,20], Yang Yu[2,20], Youyi Zhang[2,20], Hongtu Zhu[2,20], Jeffrey S. Morris[2,20], Veerabhadran Baladandayuthapani[2,20], John W. Shumway[1], Alakonanda Ghosh[1], Andrei Pöhlmann[21], Hady A. Phoulady[22], Vibhas Goyal[23], Guadalupe Canahuate[24], G. Elisabeta Marai[25], David Vock[26], Stephen Y. Lai[27], Dennis S. Mackin[15,17], Laurence E. Court[15,17], John Freymann[28], Keyvan Farahani[29,30], Jayashree Kaplathy-Cramer[31], and Clifton D. Fuller[1,2,17*] on behalf of MICCAI/M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group*

[1] Department of Radiation Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, United States, [2] Baylor College of Medicine, Houston, TX, United States, [3] Università degli Studi di Milano, Milan, Italy, [4] Department of Clinical Oncology and Nuclear Medicine, Alexandria University, Alexandria, Egypt, [5] McGovern Medical School, University of Texas, Houston, TX, United States, [6] School of Medicine, The University of Texas Health Science Center San Antonio, San Antonio, TX, United States, [7] Department of Cardiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States, [8] Furman University, Greenville, SC, United States, [9] Abilene Christian University, Abilene, TX, United States, [10] Department of Radiation Oncology, Oregon Health and Science University, Portland, OR, United States, [11] Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, United States, [12] Texas Tech University Health Sciences Center El Paso, El Paso, TX, United States, [13] University of North Texas Health Science Center, Fort Worth, TX, United States, [14] Department of Radiation Oncology, Cleveland Clinic, Cleveland, OH, United States, [15] Colgate University, Hamilton City, CA, United States, [16] Graduate School of Biomedical Sciences, MD Anderson Cancer Center, Houston, TX, United States, [17] Department of Radiation Physics, Graduate School of Biomedical Sciences, MD Anderson Cancer Center, Houston, TX, United States, [18] Moores Cancer Center, University of California, La Jolla, San Diego, CA, United States, [19] Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università Degli Studi di Napoli Federico II, Naples, Italy, [20] Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, United States, [21] Fraunhofer-Institut für Fabrikbetrieb und Automatisierung (IFF), Magdeburg, Germany, [22] Department of Computer Science, University of Southern Maine, Portland, OR, United States, [23] Indian Institute of Technology Hyderabad, Sangareddy, India, [24] University of Iowa, Iowa City, IA, United States, [25] University of Illinois at Chicago, Chicago, IL, United States, [26] Department of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, United States, [27] Department of Head and Neck Surgery, University of Texas MD Anderson Cancer Center, Houston, TX, United States, [28] Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, MD, United States, [29] National Cancer Institute, Rockville, MD, United States, [30] The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins Medicine, Baltimore, MD, United States, [31] Department of Radiology and Athinoula A. Martinos Center for Biomedical Imaging, MGH/Harvard Medical School, Boston, MA, United States

OPEN ACCESS

Edited by:
Issam El Naqa,
University of Michigan, United States

Reviewed by:
Marta Bogowicz,
UniversitätsSpital Zürich, Switzerland
Mary Feng,
University of California, San Francisco, United States

****Correspondence:***
Hesham Elhalawani
hmelhalawani@mdanderson.org
Clifton D. Fuller
cdfuller@mdanderson.org

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Radiation Oncology, a section of the journal Frontiers in Oncology

Received: *26 January 2018*
Accepted: *16 July 2018*
Published: *17 August 2018*

Citation:
Elhalawani H, Lin TA, Volpe S, Mohamed ASR, White AL, Zafereo J, Wong AJ, Berends JE, AboHashem S, Williams B, Aymard JM, Kanwar A, Perni S, Rock CD, Cooksey L, Campbell S, Yang P, Nguyen K, Ger RB, Cardenas CE, Fave XJ, Sansone C, Piantadosi G, Marrone S, Liu R, Huang C, Yu K, Li T, Yu Y, Zhang Y, Zhu H, Morris JS, Baladandayuthapani V, Shumway JW, Ghosh A, Pöhlmann A, Phoulady HA, Goyal V, Canahuate G, Marai GE, Vock D, Lai SY, Mackin DS, Court LE, Freymann J, Farahani K, Kaplathy-Cramer J and Fuller CD (2018) Machine Learning Applications in Head and Neck Radiation Oncology: Lessons From Open-Source Radiomics Challenges. Front. Oncol. 8:294. doi: 10.3389/fonc.2018.00294

Radiomics leverages existing image datasets to provide non-visible data extraction via image post-processing, with the aim of identifying prognostic, and predictive imaging features at a sub-region of interest level. However, the application of radiomics is hampered by several challenges such as lack of image acquisition/analysis method

standardization, impeding generalizability. As of yet, radiomics remains intriguing, but not clinically validated. We aimed to test the feasibility of a non-custom-constructed platform for disseminating existing large, standardized databases across institutions for promoting radiomics studies. Hence, University of Texas MD Anderson Cancer Center organized two public radiomics challenges in head and neck radiation oncology domain. This was done in conjunction with MICCAI 2016 satellite symposium using Kaggle-in-Class, a machine-learning and predictive analytics platform. We drew on clinical data matched to radiomics data derived from diagnostic contrast-enhanced computed tomography (CECT) images in a dataset of 315 patients with oropharyngeal cancer. Contestants were tasked to develop models for (i) classifying patients according to their human papillomavirus status, or (ii) predicting local tumor recurrence, following radiotherapy. Data were split into training, and test sets. Seventeen teams from various professional domains participated in one or both of the challenges. This review paper was based on the contestants' feedback; provided by 8 contestants only (47%). Six contestants (75%) incorporated extracted radiomics features into their predictive model building, either alone ($n = 5$; 62.5%), as was the case with the winner of the "HPV" challenge, or in conjunction with matched clinical attributes ($n = 2$; 25%). Only 23% of contestants, notably, including the winner of the "local recurrence" challenge, built their model relying solely on clinical data. In addition to the value of the integration of machine learning into clinical decision-making, our experience sheds light on challenges in sharing and directing existing datasets toward clinical applications of radiomics, including hyper-dimensionality of the clinical/imaging data attributes. Our experience may help guide researchers to create a framework for sharing and reuse of already published data that we believe will ultimately accelerate the pace of clinical applications of radiomics; both in challenge or clinical settings.

Keywords: machine learning, radiomics challenge, radiation oncology, head and neck, big data

INTRODUCTION

Radiomics, or texture analysis, is a rapidly growing field that extracts quantitative data from imaging scans to investigate spatial and temporal characteristics of tumors (1). To date, radiomics feature signatures have been proposed as imaging biomarkers with predictive and prognostic capabilities in several types of cancer (2–6). Nevertheless, non-uniformity in imaging acquisition parameters, volume of interest (VOI) segmentation, and radiomics feature extraction software tools make comparison between studies difficult, and highlight unmet needs in radiomics (7). Specifically, reproducibility of results is a necessary step toward validation and testing in real-world multicenter clinical trials (8). Another commonly emphasized bias of high-throughput classifiers such as those in radiomics is the "curse of dimensionality," which stems from having relatively small datasets and a massive number of possible descriptors (9).

Multi-institutional cooperation and data sharing in radiomics challenges can address, in particular, the issue of dimensionality and advance the field of quantitative imaging (10, 11). Hence, the Quantitative Imaging Network (QIN) of the National Cancer Institute (NCI) (12) started the "Challenges Task Force" with singular commitment to collaborative projects and challenges that leverage analytical assessment of imaging technologies and quantitative imaging biomarkers (13). To this end, and at the request of NCI and invitation from Medical Image Computing and Computer Assisted Intervention [MICCAI] Society, the head and neck radiation oncology group at The University of Texas MD Anderson Cancer Center organized two radiomics competitions. Oropharyngeal cancer (OPC) was chosen as a clinically relevant realm for radiomics hypothesis testing. Using manually-segmented contrast-enhanced computed tomography (CECT) images and matched clinical data, contestants were tasked with building one of 2 models. These included: (i) a classification model of human papillomavirus (HPV) status; and (ii) a predictive model of local tumor recurrence, following intensity-modulated radiation treatment (IMRT) (14).

We had several motivations for organizing these radiomics challenges. First: To demonstrate that radiomics challenges with potential clinical implementations could be undertaken for MICCAI. Second: To identify whether Kaggle in Class, a commercial educationally-oriented platform could be used as an avenue to make challenges feasible in the absence of custom-constructed websites or elaborate manpower. The main

aim of this review is to detail the mechanics and outcomes of our experience of using a large standardized database for radiomics machine-learning challenges. We previously detailed the data included in both our challenges in a recently published data descriptor (14). Here, we will continue to outline the "challenge within a challenge" to provide a template workflow for initiating substantial platforms for facilitating "multi-user" radiomics endeavors. By pinpointing these hurdles, we hope to generate insights that could be used to improve the design and execution of future radiomics challenges as well as sharing of already published radiomics data in a time-effective fashion.

MATERIALS AND METHODS FOR CHALLENGES

At the invitation of NCI and MICCAI, the head and neck radiation oncology group at The University of Texas MD Anderson Cancer Center organized two public head and neck radiomics challenges in conjunction with the MICCAI 2016: Computational Precision Medicine satellite symposium, held in Athens, Greece. Contestants with machine-learning expertise were invited to construct predictive models based on radiomics and/or clinical data from 315 OPC patients to make clinically relevant predictions in the head and neck radiation oncology sphere.

Database

After an institutional review board approval, diagnostic CECT DICOM files and matched clinical data were retrieved for OPC patients who received curative-intent IMRT at our institution between 2005 and 2012 with a minimum follow-up duration of 2 years. A key inclusion criterion was pre-treatment testing for p16 expression as a surrogate for HPV status. 315 patients with histopathologically-proven OPC were retrospectively restored from our in-house electronic medical record system, ClinicStation. The study was Health Insurance Portability and Accountability Act (HIPAA) compliant, and the pre-condition for signed informed consent was waived (15).

We then imported contrast-enhanced CT scans of intact tumor that were performed not only before the start of IMRT course but also before any significant tumor volume-changing procedures, i.e., local or systemic therapies. Although all patients were treated at the same institute, their baseline CECT scans were not necessarily obtained from the same scanner, i.e., different scanners within the same institute or less commonly baseline scans from outside institute. Hence, thorough details of images characteristics and acquisition parameters were kept in the DICOM header and made available as a **Supplementary Table**. A publicly available anonymizer toolbox, DICOM Anonymizer version 1.1.6.1, was employed to anonymize protected health information (PHI) on all DICOM files in accordance with the HIPAA, as designated by the DICOM standards from the Attribute Confidentiality Profile (DICOM PS 3.15: Appendix E) (16).

The selected CT scans were imported to VelocityAI 3.0.1 software (powered by VelocityGrid), which was used by two expert radiation oncologists to segment our VOIs in a slice-by-slice fashion. VOIs were defined as the pre-treatment gross tumor volume (GTV) of the primary disease (GTVp), which was also selected as the standardized nomenclature term. Gross nodal tumor volumes also were segmented to provide a complete imaging dataset that can benefit other radiomics studies in the head and neck cancer domain. However, contestants were clearly instructed to include only GTVp in regions of interest for robust texture analysis.

Segmented structures in congruence with matched clinical data constituted the predictor variables for both challenges. Clinical data elements comprised patient, disease, and treatment attributes that are of established prognostic value for OPC (17). A matching data dictionary of concise definitions, along with possible levels for each clinical data attribute, was provided to contestants as a "ReadMe" CSV file (**Table 1**).

We also provided contestants with a list of suggested open-source infrastructure software that supports common radiomics workflow tasks such as image data import and review as well as radiomics feature computation, along with links to download the software. After completion of the challenge, a complete digital repository was deposited (figshare: https://doi.org/10.6084/m9.figshare.c.3757403.v1 and https://doi.org/10.6084/m9.figshare.c.3757385.v1) (18, 19) and registered as a public access data descriptor (14).

Challenge Components

Challenge components were identified as a function of the hosting platform.

Hosting Platform

In the two radiomics challenges, organized on Kaggle-in-Class, contestants were directed to construct predictive models that (i) most accurately classified patients as HPV positive or negative compared with their histopathologic classification (http://inclass.kaggle.com/c/oropharynx-radiomics-hpv), and (ii) best predicted local tumor recurrence (https://inclass.kaggle.com/c/opc-recurrence). Kaggle-in-Class (https://inclass.kaggle.com/) is a cloud-based platform for predictive modeling and analytics contests on which researchers post their data and data miners worldwide attempt to develop the most optimal predictive models. The overall challenge workflow is portrayed in **Figure 1**.

Anonymized imaging and clinical data belonging to the cohort of 315 OPC patients were uploaded to the Kaggle in Class server almost evenly split between the training subset and test subset, encompassing 150 and 165 patients, respectively, in separate CSV files and DICOM folders. Subjects were randomly assigned to either training or test sets via random number generation. Caution was taken to make outcome of interest (HPV status for the first challenge and local control for the second one)

TABLE 1 | Supplemental information about data provided for radiomics challenges.

Data element	Description
Patient ID	Numbers given randomly to the patient after anonymization of the DICOM protected health identifier (PHI) tag (0010,0020) that corresponds to medical record number
HPV/p16 status	HPV status, as assessed by HPV DNA *in situ* hybridization (57) and/or p16 protein expression via immunohistochemistry, with the results described as 1 (i.e., positive) or 0 (i.e., negative)
Gender	Patient's sex
Age at diagnosis	Patient's age in years at the time of diagnosis
Race	American Indian/Alaska Native, Asian, Black, Hispanic, White, or not applicable
Tumor laterality	Right, left, or bilateral
Oropharynx subsite of origin	Subsite of the tumor within the oropharynx, i.e., base of tongue (21) or tonsil/soft palate/pharyngeal wall/glossopharyngeal sulcus/other (no single subsite of origin could be identified)
T category	Description of the original (primary) tumor with regard to size and extent per the American Joint Committee on Cancer (AJCC) and Union for International Cancer Control (UICC) cancer staging system, i.e., T1, T2, T3, or T4 (https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx)
N category	Description of whether the cancer has reached nearby lymph nodes, per the AJCC and UICC cancer staging system, i.e., N0, N1, N2a, N2b, N2c, or N3 (https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx)
AJCC stage	AJCC cancer stage (https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx)
Pathologic grade	Grade of tumor differentiation, i.e., I, II, III, IV, I-II, II-III, or not assessable
Smoking status at diagnosis	Never, current, or former smoker
Smoking pack-years	An equivalent numerical value of lifetime tobacco exposure; 1 pack-year is defined as 20 cigarettes smoked every day for 1 year

proportionally distributed in training and test sets. For the test set, contestants were blinded to the outcome.

Evaluation Metric

The evaluation metric for both competitions was area under the receiver operating characteristic curve (AUC) of the binary outcomes, i.e., "positive" vs. "negative" for the "HPV" challenge or "recurrence" vs. "no recurrence" for the "local recurrence" challenge.

Scoring System

Kaggle-in-Class further splits the test set randomly into two subsets of approximately equal size again with outcome of interest equally distributed. One subset was made public to contestants, named the "Public Test subset." The other subset was held out from the contestants, with only challenge organizers having access to it, named the "Private Test subset." The performance of the contestants' models was first assessed on the public test set and results were posted to a "Public leaderboard." The public leaderboards were updated continuously as contestants made new submissions, providing real-time feedback to contestants on the performance of their models on the public test subset relative to that of other contestants' models.

The private leaderboard was accessible only to the organizers of the challenges. Toward the end of the challenge, each contestant/team was allowed to select his/her/their own two "optimal" final submissions of choice. Contestants were then judged according to the performance of their chosen model(s) on the private test subset, according to the private leaderboard. The contestant/team that topped the "private leaderboard" for each challenge was declared the winner of the challenge. The distinction between training/test set and public/private subset terminology is further illustrated in **Figure 2**.

Challenges Rules

Teams were limited to a maximum of two result submissions per team per day. There was no maximum team size, but merging with or privately sharing code and data with other teams was prohibited.

Challenges Organizers-Contestants Interaction

To enable contestants to communicate with the organizing committee, the e-mail address of one of the organizers was made available on the Kaggle in Class and MICCAI websites. Also, the organizers created and closely followed a discussion board where updates or topics of common interest were publicly shared. After announcing the winners, questionnaires were distributed to contestants to get their feedback, which greatly contributed to this review paper.

CHALLENGE RESULTS

Seventeen teams participated in either one or both challenges, accounting for a total of 23 enrollments. The "HPV" challenge recorded nine enrollments comprising three multiple-member teams and six individual contestants. The "local recurrence" challenge, on the other hand, had four multiple-member teams and 10 individual contestants. The following results are derived from the questionnaires, which were filled out by eight teams. Detailed responses of contestants to post-challenges surveys are tabulated in **Supplementary Table 1**. Contestants came from various professional domains, e.g., biostatistics, computer science, engineering, medical physics, mathematics, and radiation oncology. The dedicated time per participant for each challenge ranged from 6 to 30 h. Teams included as many as seven members with the same or different institutional affiliations.

The data analytical algorithms showed wide variation in methods and implementation strategies. The programming

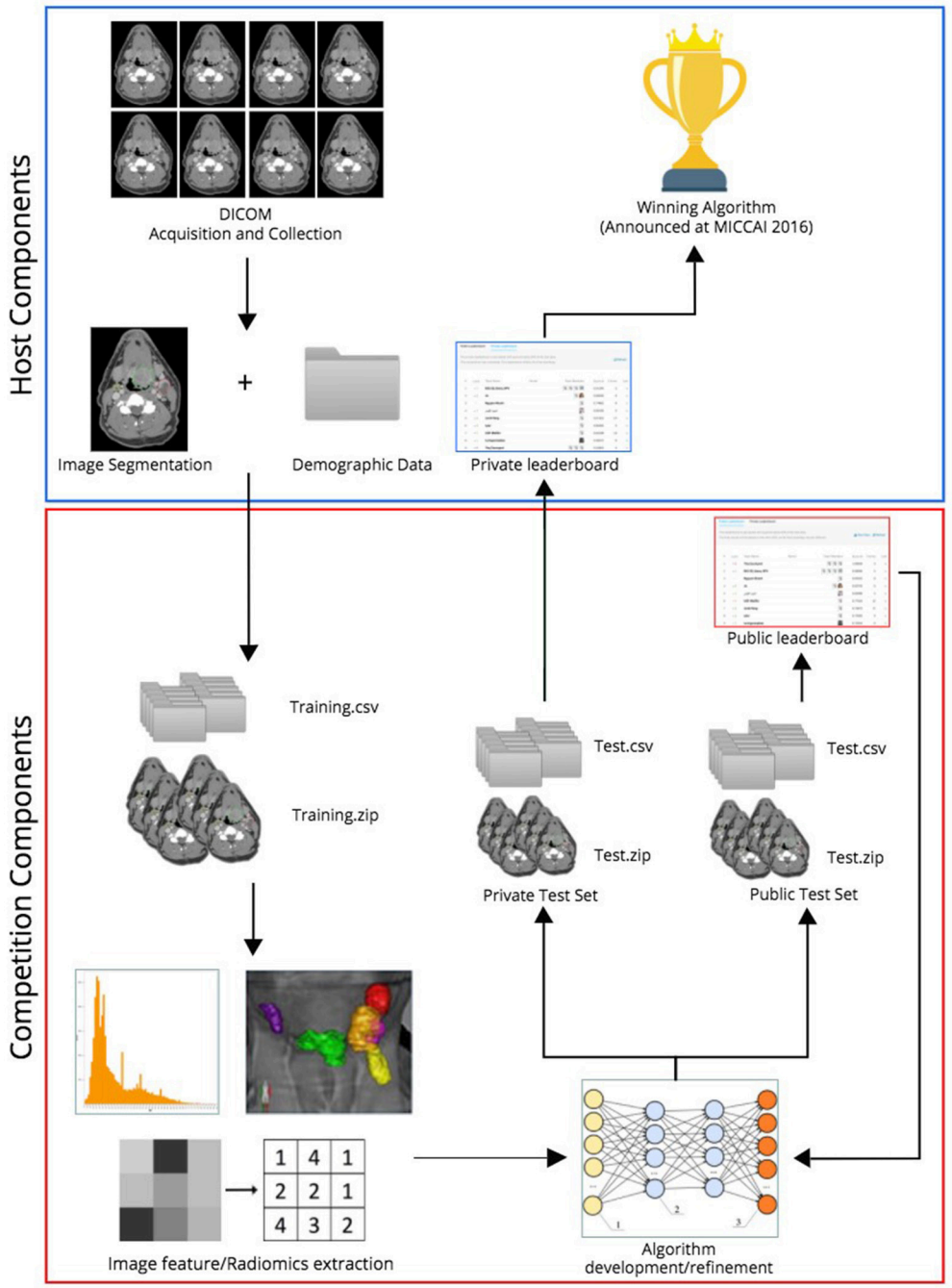

FIGURE 1 | Workflow of radiomics challenges.

platforms used to extract quantitative radiomics features included MATLAB, R, and Python. Most contestants (63%) developed their own scripts to extract radiomics features. The Imaging Biomarker Explorer (IBEX) software, developed by the Department of Radiation Physics at MD Anderson (20), was the second most commonly used software among the other contestants (38%). The machine-learning techniques used included random forest with class balancing, logistic regression with gradient descent or extreme gradient boosting trees, least absolute shrinkage, and selection operator (Lasso) regression, and neural networks. Interestingly, one contestant reported applying an ensemble combination of classifiers, including random forests, a naïve Bayes classifier, and Association

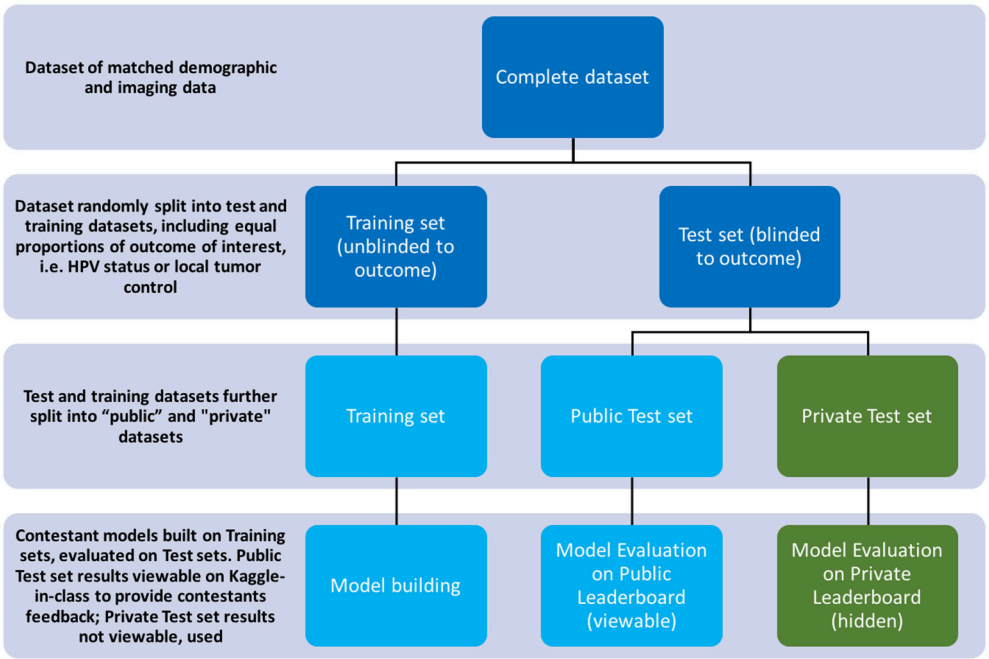

FIGURE 2 | Diagram illustrating the splitting of datasets per the challenge's rules.

for Computing Machinery classifiers, as well as boosting algorithms, including AdaBoost, and oversampling techniques, including Synthetic Minority Over-sampling Technique. The most commonly used statistical tests included leave-one-out cross-validation, the Wilcoxon rank-sum test, and sparse matrices.

The key, relevant radiomics features selected by these various machine-learning algorithms encompassed various first- and second-order features. The chosen first-order features included the "intensity" feature of maximum intensity and the "shape" features of primary tumor volume, longest and shortest radii, and Euclidean distance (in mm, with respect to centroids) between the primary tumor and the lymph nodes (minimum, maximum, mean, and standard deviation). The chosen second-order features included gray-level co-occurrence matrix and local binary pattern.

Key clinical data commonly selected and modeled by contestants included smoking pack-years, T category, N category, and tumor subsite of origin, e.g., tonsil or base of tongue. Most contestants (77%) incorporated extracted radiomics features into their model, either alone (62%), as was the case with the winning team of the "HPV" challenge, or in conjunction with matched clinical attributes (16%). Meanwhile, only 23% of contestants built their models relying solely on clinical data, including the winner of the "local recurrence" challenge.

Per contestant feedback, the obstacles to developing sound machine-learning predictive models were largely technical in nature. Fifty percent of questionnaire respondents reported inability to extract radiomics features, especially global directional features, for some images. This was the leading cause of missing values, which were difficult to handle for most contestants. Other barriers involved segmentation issues where some VOIs—according to one contestant—were not consistently named across the whole cohort. A few contestants also reported that some GTVp contours did not adequately represent the primary tumor lesions, i.e., some slices within the VOI were not segmented, or GTVp contours were totally absent. In some cases, only metastatic lymph nodes (i.e., gross nodal tumor volume) were segmented, per one contestant. Nonetheless, all but one team expressed enthusiasm toward participating in future machine-learning challenges.

For the "HPV" challenge, the winners were a team of academic biostatisticians with a radiomics-only model that achieved an AUC of 0.92 in the held-out, private test subset. Their feature selection approach yielded the "shape" features of "mean breadth" and "spherical disproportion" as most predictive of HPV status, suggesting that HPV-associated tumors tend to be smaller and more homogeneous. On the other hand, the winner of the "local recurrence" challenge was a mathematics/statistics college student who exclusively used clinical features to build a model that achieved an AUC of 0.92 in the private test subset. The AUCs of all contestants' models and their corresponding final ranking in the private leaderboard are provided in **Supplementary Tables 2**, **3**.

The winner of each challenge was invited to share their approach and models via video conference at the Computational Precision Medicine satellite workshop as part of the MICCAI 2016 program that took place in Athens, Greece. Moreover, each winner was offered a manuscript acceptance (after editorial

review) with fees waived to describe their approach and algorithm in an international, open-access, peer-reviewed journal sponsored by the European Society for Radiotherapy and Oncology. The winners of the "HPV" challenge recently reported their approach in designing a statistical framework to analyze CT images to predict HPV status (21).

DISCUSSION

The process of designing and executing the radiomics challenge was inevitably filled with difficult decisions and unexpected issues, from which we have yielded numerous insights. We have enumerated these challenges and derived lessons in **Table 2**.

"Challenge Within a Challenge" and Derived Lessons

Before, during and even after the radiomics challenges, we encountered situations which provided us insight into improving future radiomics challenges. We will now detail learning points derived from our experience.

Database Size

The usefulness of a database for radiomics analysis increases as more and more cases are added. However, limits on time, personnel, and available patient data place constraints on database collection and thus the ability to yield insights from radiomics analysis. Also relevant to imaging data collection is the variation in imaging acquisition parameters and disease states within a disease cohort. In our case, as in many practical classification problems, HPV status and local control rates following IMRT for OPC patients tend to be imbalanced. The majority of OPC tend to be increasingly associated with HPV infection and hence more favorable local control (22). In our dataset, HPV-negative and locally recurrent OPC only constituted 14.9 and 7.6% of the overall cohort, respectively.

Moreover, the enormous number of potential predictor variables used in radiomics studies necessitates the use of large-scale datasets in order to overcome barriers to statistical inference (23). The dearth of such datasets hinders machine-learning innovation in radiation oncology by restricting the pool of innovation to the few institutions with the patient volume to generate usable datasets (24).

TABLE 2 | Challenges and derived lessons from organizing open-source radiomics challenges.

Challenges
• Paucity of open-source freely available radiomics datasets
• Establishing database: size vs. time
• Data anonymization
• Quality assurance: before, during, and after the challenge
• Understanding contestants' preferences
• Clarity of challenge rules verbiage
• Hyperdimensionality of radiomics co-variates and subsequent overtraining
• Low post-challenge survey response rate
• Discrete scanners, acquisition parameters, and segmentation techniques
Derived Lessons
• Use common ontology guidelines to assign nomenclature for target volumes and clinical data
• Use efficient, secure solution such as RSNA CTP to minimize time/resource burden
• Test run data prior to start of radiomics challenge to identify additional issues
• Adopt "Public/Private leaderboard" challenges to mitigate overtraining/overfitting
• Choice of data type and sources (i.e., single vs. multi-institutional) depends on specific aims of radiomics challenge
• Provide contestants multiple ways to analyze data whenever possible, e.g., with/without artifacts to account for variation in contestants' preferences
• Rules must be clear and consistent with all other aspects of challenge design
• Proper incentives built into the radiomics challenge encourage participation and subsequent feedback
• Post-challenge permanent data repository and descriptor

Data Anonymization

The PHI anonymization software we applied was cumbersome, requiring PHI tags to be manually entered on an individual basis. For future radiomics challenges, we recommend the use of the Clinical Trial Processor (CTP), developed by the Radiological Society of North America (RSNA) (25). Safe, efficient, and compatible with all commercially available picture archiving and communication systems (PACS), RSNA CTP is designed to transport images to online data repositories (25). RSNA CTP conforms closely to image anonymization regulations per the HIPAA Privacy Rule and the DICOM Working Group 18 Supplement 142 (16).

Data Curation and Standardization

Standardization and harmonization of data attributes provide the foundation for developing comparable data among registries that can then be combined for multi-institutional studies (26). This further empowers validation studies and subsequent generalization of the resulting models from such studies. In our challenges, VOIs were not consistently coded across the whole cohort, according to one contestant, a finding necessitating our correction to facilitate subsequent analysis for contestants.

Hence, we recommend conforming to common ontology guidelines when assigning nomenclature for target volumes and clinical data. Good examples would be the American Association of Physicists in America Task Group 263 (AAPM TG-263) (27) and North American Association of Central Cancer Registries (NAACCR) guidelines (28).

Volume of Interest Definition and Delineation

Another cumbersome aspect of data curation is the segmentation of target volumes. Reliable semi-automated segmentation methods for head and neck carcinomas and normal tissues are currently still under investigation, so we relied on manual segmentation (29, 30). The disadvantages of manual segmentation relate not only to being time-consuming but also to intra- and inter-observer variability (31). A collateral benefit of making CT datasets with expert manual segmentations publicly available is testing semi-automated segmentation tools (32).

In our case, 2 radiation oncologists were blinded to relevant clinical data and outcomes, and their segmentations were cross-checked then double-checked by a single expert radiation oncologist, to diminish inter-observer variability. Guidelines of the International Commission on Radiation Units and Measurements reports 50 and 62 were followed when defining target volumes (33, 34).

Scanner and Imaging Parameters Variability

Variability in inter-scanner and imaging acquisition parameters, like voxel size, reconstruction kernel, tube current and voltage has been shown to influence radiomics analyses (35–39). Thus, when sharing imaging data with contestants and uploading to public data repositories, we recommend preserving all DICOM headers aside from those containing protected health information. These parameters, easily extractable from DICOM headers, can also be provided as **Supplementary Materials** for future radiomics challenges. Although we did not elicit specific feedback in the post-challenge survey regarding how contestants accounted for differences in image acquisition, we recognize the importance of this question and recommend its inclusion in future radiomics challenge contestant surveys.

Moreover, head and neck radiomics are subject to the effects of image artifacts from intrinsic patient factors, such as metal dental implants and bone. The effects of resulting streak artifacts and beam-hardening artifacts on robustness of extracted radiomics features have been reported (3, 40). Our approach within this study was to remove slices of the GTV on computed tomography that were affected by artifacts. However, this results in missing information or contours that do not adequately represent the primary tumor lesion, as was noted by some contestants.

Single-institutional radiomics databases like the one used in our challenges minimize inter-scanner variability. However, in some cases the increased heterogeneity of multi-institutional databases is preferred. The choice (i.e., single vs. multi-institutional data) should be challenge-dependent. Single-institutional data may be preferred if uniformity in some imaging characteristics (e.g., slice thickness, acquisition protocol) is required for exploratory research purposes. Multi-institutional data are preferred as the end goal of radiomics challenges and studies is to generate clinically relevant models with maximum generalizability to other patient populations.

Interplay Between Clinical and Radiomics Data Variables

We sought to provide the option to include not only physical variables but also key clinical attributes in the model building. We aimed to test the capacity of radiomics features, alone or in combination with clinical features, to model classification or risk prediction scenarios. Interestingly, the winner of the "local recurrence" challenge and the winner of the "HPV" challenge used only clinical and only radiomics data, respectively. Ironically, the fact that some contestants could generate more effective non-radiomics models for risk prediction may subvert the entire aim of the challenge. This in turn demonstrates the difficulty of integrating radiomics into clinical data in both challenge and clinical settings.

In the OPC setting, we recommend that HPV status be provided for all cases, being an independent prognostic and predictive biomarker in the OPC disease process (17, 41). However, it is also important for future radiomics challenges to consider whether other clinically relevant factors like smoking history, tumor subsite, or race are pertinent to the end goal of their challenge.

Quality Assurance

It is important for quality assurance measures used in radiomics challenges to mirror those of traditional radiomics studies. If the dataset has not been used in a radiomics analyses, it is imperative for test analyses to identify errors. Although we had quality assurance protocols in place, contestants still noted issues with the dataset. Using Kaggle in Class, contestants were able to report feedback in real time. In turn, the responses we posted to the Kaggle in Class Forum could be viewed by all groups, ensuring that all contestants had access to the same updated information at all times, regardless of who originally asked a question. As the challenge progressed, contestants reported 9 corrupt, inaccessible DICOM imaging files and 18 patients with GTVps which did not adequately encompass the primary gross tumor volume. In other cases, the GTVp contours were absent, meaning these patients only had GTVn contours—the use of which was prohibited by challenges rules. Although we responded to contestant feedback in real time, we believe that clear and explicitly stated challenge rules as well as an initial test run of the data are essential.

Recruiting Contestants

Participation in the radiomics challenges by academic groups with radiomics expertise was lower than anticipated. This reticence may be due to the public nature of the challenge combined with the uncertainty of success inherent in analyzing new datasets in limited timeframes, as well as the lack of clear translation to publishable output. An alternative explanation is that machine learning challenges platforms like Kaggle in Class are less well known to the radiomics community in comparison to the MICCAI community.

To attract contestants with radiomics expertise, it is necessary to ensure proper incentives are in place. Challenge announcements should be made well in advance of the challenge start date to provide sufficient time for contestants to include the challenge into their work plans. Partnering with renown organizations like NCI QIN and MICCAI on the challenge provides institutional branding which may draw in academic groups. Offers of co-authorship on future publications stemming from the challenge, as well as seats on conference panels at which challenge results will be shared, may boost participation.

Email distribution lists of professional societies such as MICCAI, SPIE (The International Society for optics and photonics) Medical Imaging, and The Cancer Imaging Archive (TCIA) would be an effective way to reach academics. Platforms like Kaggle and KDnuggets are more popular among non-academics interested in machine learning challenges.

Understanding Contestant Preferences

Contestants in our challenge wished to have additional data beyond what was provided. For instance, multiple contestants noted that some patients had missing VOIs on certain slices of the image. We had made the choice to omit these slices because the VOI in these regions was significantly obscured by dental artifact. However, contestants felt that shape and spatially-derived features might be affected by omission of these slices. To avoid this situation in future radiomics challenges, we suggest providing two datasets, one with artifacts included and one with artifacts excluded. This arrangement allows contestants the choice of which dataset to analyze.

Public and Private Leaderboards

The problem of overfitting has been observed in previous radiomics studies (7). Blinding contestants to their model's performance on the private test subset ensured that contestants were not overfitting their data to the test set. Hence, we chose the Kaggle in Class platform to host the challenges because it offers both public and private leaderboards based on public and held-out subsets of the test dataset, respectively. This design choice appeared to serve its intended purpose. In the "HPV" challenge, the first-place team on the public leaderboard had an AUC of 1.0 but finished in last place on the private leaderboard with an AUC of 0.52. This discrepancy suggests that their model suffered from overfitting issues. In contrast, the winner of the "HPV" challenge performed well on both public and private leaderboards, indicating that their proposed model was more generalizable.

Clarity of Challenges Rules

One difficulty inherent in radiomics challenges is variability in interpretation of challenge rules. This variability may be driven by differences in contestants' technical expertise, culture, background, and experiences. Thus, clear and unambiguous rules and challenge design are desirable. For example, our challenge rules clearly stated that radiomics features should be exclusively extracted from GTVp. However, GTVp was unavailable for some patients, typically post-surgical patients with no available pre-treatment imaging. When combined with the fact that we also provided GTVn for all patients, some contestants were confused by the conflicting messages they received. Thus, to prevent confusion it is important that the stated rules of the challenge be consistent with all other aspects of the contestants' experience during the challenge.

Furthermore, while the challenges were branded as "radiomics challenges," we allowed the submission of models based solely on clinical prognostic factors, as was the case for the winner of the "local recurrence" challenge. In some instances, a clinical-only model may be useful as a comparison tool to determine whether there is an incremental benefit to leveraging radiomics data compared to clinical-only models. However, the permissibility of clinical-only models in radiomics challenges must be stated explicitly in contest rules to prevent confusion.

Collecting Contestants' Feedback

Another learning point relates to increasing post-contest survey response rates. A mere 50% of contestants responded to our post-challenge survey. To ensure a high survey response rate, we suggest including a pre-challenge agreement in which contestants pledge to complete the post-challenge survey as part of the challenge. A manuscript co-authorship contingent upon survey participation might also incentivize more contestants to fill out the survey.

Contestants' Responsibilities

Participation in radiomics challenges necessitates a good-faith commitment on the part of contestants to follow through with the challenge, even in the face of unsatisfactory model performance. Withdrawals are antithetical to the mission of radiomics challenges as a learning tool for both challenge contestants and organizers to advance the field.

Permanent Data Repositories

The decision to upload our dataset to an online data repository, in this case figshare (https://doi.org/10.6084/m9.figshare.c.3757403.v1 and https://doi.org/10.6084/m9.figshare.c.3757385.v1) (18, 19), was not difficult. This was done to provide a curated OPC database for future radiomics validation studies. Furthermore, all contestants who downloaded the database during the challenge would already have access to the data, and it would have been impractical to ask all contestants to delete this information once downloaded.

We are also in the process of uploading this dataset as a part of a larger matched clinical/imaging dataset to TCIA. Versioning, which is a built-in feature in most data repositories including figshare, is essential for updating datasets, e.g., following quality assurance as well as retrieving previous versions later. To date, we have received multiple requests to use our dataset for external validation of pre-existing models.

We chose not to make available the "ground truth" of the private test subset data. The decision to withhold this information diminishes the overall value of the database to researchers using the dataset but in return preserves these test cases for future challenges.

Post-challenge Methodology and Results Dissemination

One potential obstacle to disseminating radiomics challenge results relates to participant requests for anonymity. A participant's right, or lack thereof, to remain anonymous in subsequent publications of challenge results must be stated prior to the start of the challenge. Anonymity poses issues with reporting methodologies and subsequent model performance results, as these results may be traceable to the original online Kaggle in Class challenge website, where identities are not necessarily obscured. Transparency of identities, methodologies, and results is in the spirit of data sharing and is our preferred arrangement in radiomics challenges.

Scientific papers analyzing the individual performances of winning algorithms submitted to the Challenge, along with database descriptor have been or will be published (14, 21). In general, we also recommend publishing a post-challenge data descriptor that details data configuration as a guide for future dataset usage (14).

Conclusions and Future Outlook

In summary, the MICCAI 2016 radiomics challenges yielded valuable insights into the potential for radiomics to be used in clinically relevant prediction and classification questions in OPC. Furthermore, our experience designing and executing the radiomics challenge imparted lessons which we hope can be applied to the organization of future radiomics challenges, such as those associated with the MICCAI 2018 Conference.

DATA AVAILABILITY STATEMENT

Datasets are in a publicly accessible repository: The datasets generated for this study can be found in figshare; https://doi.org/10.6084/m9.figshare.c.3757403.v1 and https://doi.org/10.6084/m9.figshare.c.3757385.v1.

AUTHOR CONTRIBUTIONS

Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; Drafting the work or revising it critically for important intellectual content; Final approval of the version to be published; Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Specific additional individual cooperative effort contributions to study/manuscript design/execution/interpretation, in addition to all criteria above are listed as follows: HE manuscript writing, direct oversight of all image segmentation, clinical data workflows, direct oversight of trainee personnel (ALW, JZ, AJW, JB, SA, BW, JA, and SP). TAL, SV, and PY wrote sections of the manuscript. AM primary investigator; conceived, coordinated, and directed all study activities, responsible for data collection, project integrity, manuscript content and editorial oversight and correspondence. AK, ALW, JZ, AJW, JB, SC, and SP clinical data curation, data transfer, supervised statistical analysis, graphic construction, supervision of DICOM-RT analytic workflows and initial contouring. SA, BW, JA, and LC electronic medical record screening, automated case identification, data extraction, clinical. Participated in at least one radiomics challenge, submitted a valid results and completed post-challenge questionnaire (KN, RG, CC, XF, CS, GP, SM, RL, CH, KY, TL, VB, JS, AG, AP, HP, VG, GC, GM, DV, SL, DM, LEC, JF, KF, JK, CF).

FUNDING

Multiple funders/agencies contributed to personnel salaries or project support during the manuscript preparation interval. Dr. HE is supported in part by the philanthropic donations from the Family of Paul W. Beach to Dr. G. Brandon Gunn, MD. This research was supported by the Andrew Sabin Family Foundation; Dr. CF is a Sabin Family Foundation Fellow. Drs. SL, AM, and CF receive funding support from the National Institutes of Health (NIH)/National Institute for Dental and Craniofacial Research (1R01DE025248-01/R56DE025248-01). Drs. GM, DV, GC, and CF are supported via a National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679). Dr. CF received grant and/or salary support from the NIH/National Cancer Institute (NCI) Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50 CA097007-10) and the Paul Calabresi Clinical Oncology Program Award (K12 CA088084-06), the Center for Radiation Oncology Research (CROR) at MD Anderson Cancer Center Seed Grant; and the MD Anderson Institutional Research Grant (IRG) Program. Dr. JK-C is supported by the National Cancer Institute (U24 CA180927-03, U01 CA154601-06. Mr. Kanwar was supported by a 2016–2017 Radiological Society of North America Education and Research Foundation Research Medical Student Grant Award (RSNA RMS1618). GM's work is partially supported by National Institutes of Health (NIH) awards NCI-R01-CA214825, NCI-R01CA225190, and NLM-R01LM012527; by National Science Foundation (NSF) award CNS-1625941 and by The Joseph and Bessie Feinberg Foundation. Dr. CF received a General Electric Healthcare/MD Anderson Center for Advanced Biomedical Imaging In-Kind Award and an Elekta AB/MD Anderson Department of Radiation Oncology Seed Grant. Dr. CF has also received speaker travel funding from Elekta AB. None of these industrial partners' equipment was directly used or experimented with in the present work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2018.00294/full#supplementary-material

REFERENCES

1. Wong AJ, Kanwar A, Mohamed AS, Fuller CD. Radiomics in head and neck cancer: from exploration to application. *Transl Cancer Res.* (2016) 5:371–82. doi: 10.21037/tcr.2016.07.18
2. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. *Radiology* (2016) 281:947–57. doi: 10.1148/radiol.2016152234
3. Leijenaar RT, Carvalho S, Hoebers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol.* (2015) 54:1423–9. doi: 10.3109/0284186X.2015.1061214
4. Liang C, Huang Y, He L, Chen X, Ma Z, Dong D, et al. The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. *Oncotarget* (2016) 7:31401–12. doi: 10.18632/oncotarget.8919
5. Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol.* (2015) 60:5471–96. doi: 10.1088/0031-9155/60/14/5471
6. Elhalawani H, Kanwar A, Mohamed ASR, White A, Zafereo J, Fuller CD, et al. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. *Sci Rep.* (2018) 8:1524. doi: 10.1038/s41598-017-14687-0

7. Limkin EJ, Sun R, Dercle L, Zacharaki EI, Robert C, Reuze S, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Annals* of oncology : official journal of the European Society for Medical Oncology. (2017) 28:1191–206. doi: 10.1093/annonc/mdx034
8. O'Connor JPB, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. (2017) 14:169–86. doi: 10.1038/nrclinonc.2016.162
9. Zimek A, Schubert E, Kriegel H-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat Anal Data Mining* (2012) 5:363–87. doi: 10.1002/sam.11161
10. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology* (2011) 258:906–14. doi: 10.1148/radiol.10100799
11. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. *Nat Rev Clin Oncol*. (2013) 10:27–40. doi: 10.1038/nrclinonc.2012.196
12. Farahani K, Kalpathy-Cramer J, Chenevert TL, Rubin DL, Sunderland JJ, Nordstrom RJ, et al. Computational challenges and collaborative projects in the nci quantitative imaging network. *Tomography* (2016) 2:242–9. doi: 10.18383/j.tom.2016.00265
13. Armato SG, Hadjiiski LM, Tourassi GD, Drukker K, Giger ML, Li F, et al. LUNGx challenge for computerized lung nodule classification: reflections and lessons learned. *J Med Imaging* (2015) 2:020103. doi: 10.1117/1.JMI.2.2.020103
14. Elhalawani H, White AL, Zafereo J, Wong AJ, Berends JE, AboHashem S, et al. Fuller. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci Data* (2017) 4:170077. doi: 10.1038/sdata.2017.77
15. Freymann JB, Kirby JS, Perry JH, Clunie DA, Jaffe CC. Image data sharing for biomedical research—meeting HIPAA requirements for de-identification. *J Digital Imaging* (2012) 25:14–24. doi: 10.1007/s10278-011-9422-x
16. Fetzer DT, West OC. The HIPAA privacy rule and protected health information: implications in research involving DICOM image databases. *Acad Radiol*. (2008) 15:390–5. doi: 10.1016/j.acra.2007.11.008
17. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tân PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *New Engl J Med*. (2010) 363:24–35. doi: 10.1056/NEJMoa0912217
18. Clifton F, Abdallah M, Hesham E. *Predict from CT data the HPV phenotype of oropharynx tumors; compared to ground-truth results previously obtained by p16 or HPV testing*. *Figshare* (2017) 22:26. doi: 10.6084/m9.figshare.c.3757403.v1
19. Fuller C, Mohamed A, Elhalawani H. Determine from CT data whether a tumor will be controlled by definitive radiation therapy. *Figshare* (2017) doi: 10.6084/m9.figshare.c.3757385.v1
20. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. *Med Phys*. (2015) 42:1341–53. doi: 10.1118/1.4908210
21. Yu K, Zhang Y, Yu Y, Huang C, Liu R, Li T, et al. Radiomic analysis in prediction of Human Papilloma Virus status. *Clin Transl Radiat Oncol*. (2017) 7:49–54. doi: 10.1016/j.ctro.2017.10.001
22. Mehanna H, Beech T, Nicholson T, El-Hariry I, McConkey C, Paleri V, et al. Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer—systematic review and meta-analysis of trends by time and region. *Head Neck* (2013) 35:747–55. doi: 10.1002/hed.22015
23. Pekalska E, Duin RPW. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. Hackensack, NJ: World Scientific Publishing Co., Inc. (2005).
24. Mayo CS, Kessler ML, Eisbruch A, Weyburne G, Feng M, Hayman JA, et al. The big data effort in radiation oncology: data mining or data farming? *Adv Radiat Oncol*. (2016) 1:260–71. doi: 10.1016/j.adro.2016.10.001
25. Radiological Society of North America I. *CTP-The RSNA Clinical Trial Processor. Radiological Society of North America, Inc*. Available online at: http://mircwiki.rsna.org/index.php?title=CTP-The_RSNA_Clinical_Trial_Processor (Accessed December 1 2017).
26. Mayo CS, Pisansky TM, Petersen IA, Yan ES, Davis BJ, Stafford SL, et al. Establishment of practice standards in nomenclature and prescription to enable construction of software and databases for knowledge-based practice review. *Pract Radiat Oncol*. (2016) 6:e117–26. doi: 10.1016/j.prro.2015.11.001
27. Mayo CS, Moran JM, Bosch W, Xiao Y, McNutt T, Popple R, et al. American Association of Physicists in Medicine Task Group 263: standardizing nomenclatures in radiation oncology. *Int J Radiat Oncol Biol Phys.* (2018) 100:1057–66. doi: 10.1016/j.ijrobp.2017.12.013
28. Hulstrom DE. *Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Seventh Edition, Version 10*. Springfield, IL: North American Association of Central Cancer Registries (2002).
29. Ibragimov B, Korez R, Likar B, Pernuš F, Xing L, Vrtovec T. Segmentation of pathological structures by landmark-assisted deformable models. *IEEE Transac Med Imaging* (2017) 36:1457–69. doi: 10.1109/TMI.2017.2667578
30. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys. 2017;44:547–57. doi: 10.1002/mp.12045
31. Wu J, Tha KK, Xing L, Li R. Radiomics and radiogenomics for precision radiotherapy. *J Radiat Res*. (2018) 59(Suppl. 1):i25–31. doi: 10.1093/jrr/rrx102
32. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS ONE* (2014) 9:e102107. doi: 10.1371/journal.pone.0102107
33. ICRU Report 50. *Prescribing, Recording, and Reporting Photon Beam Therapy ICRU*. Bethesda; Oxford: Oxford University Press (1993).
34. ICRU Report 62. *Prescribing, Recording, and Reporting Photon Beam Therapy (Supplement to ICRU Report 50)ICRU*. Bethesda; Oxford: Oxford University Press (1999).
35. Fave X, Cook M, Frederick A, Zhang L, Yang J, Fried D, et al. Preliminary investigation into sources of uncertainty in quantitative imaging features. *Comput Med Imaging Graph.* (2015) 44:54–61. doi: 10.1016/j.compmedimag.2015.04.006
36. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring CT scanner variability of radiomics features. *Invest Radiol*. (2015) 50:757–65. doi: 10.1097/RLI.0000000000000180
37. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLoS ONE* (2017) 12:e0178524. doi: 10.1371/journal.pone.0178524
38. Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. (2017) 44:1050–62. doi: 10.1002/mp.12123
39. Mackin D, Ger R, Dodge C, Fave X, Chi P-C, Zhang L, et al. Effect of tube current on computed tomography radiomic features. *Sci Rep*. (2018) 8:2354. doi: 10.1038/s41598-018-20713-6
40. Block AM, Cozzi F, Patel R, Surucu M, Hurst N Jr., Emami B, et al. Radiomics in head and neck radiation therapy: impact of metal artifact reduction. *Int J Radiat Oncol Biol Phys*. (2017) 99:E640. doi: 10.1016/j.ijrobp.2017 06.2146
41. Rosenthal DI, Harari PM, Giralt J, Bell D, Raben D, Liu J, et al. Association of human papillomavirus and p16 status with outcomes in the IMCL-9815 phase III registration trial for patients with locoregionally advanced oropharyngeal squamous cell carcinoma of the head and neck treated with radiotherapy with or without cetuximab. *J Clin Oncol*. (2016) 34:1300–8. doi: 10.1200/JCO.2015.62.5970

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership