Deep learning for radiotherapy outcome prediction using dose data – a review

Abstract

Artificial intelligence (AI), and in particular deep learning using convolutional neural networks (CNN), has been used extensively for image classification and segmentation, including on medical images for diagnosis and prognosis prediction. Use in radiotherapy prognostic modelling is still limited, especially as applied to toxicity and tumour response prediction from radiation dose distributions. We review and summarise studies which apply deep learning to radiotherapy dose data, in particular studies which utilise full 3D dose distributions. Ten papers have reported on deep learning models for outcome prediction utilising spatial dose information while four studies used reduced dimensionality (DVH, Dose Volume Histogram) information for prediction. Many of these studies suffer from some of the same issues which plagued early normal tissue complication probability (NTCP) modelling; including small, single-institutional patient cohorts, lack of external validation, poor data and model reporting, use of late toxicity data without taking time-to-event into account, and nearly exclusive focus on clinician-reported complications. They demonstrate, however, how radiation dose, imaging, and clinical data may be technically integrated in CNN-based models; and some studies explore how deep learning may help better understand spatial variation in radiosensitivity. In general, there are a number of issues specific to the intersection of radiotherapy outcome modelling and deep learning, for example translation of model developments into treatment plan optimisation, which will require further combined effort from the radiation oncology and AI communities.

Keywords

Artificial intelligence, deep learning, outcome prediction, radiotherapy

Statement of search strategies used

For this scoping review, we conducted a systematic search of MEDLINE via PubMed, from inception to August 2021. We searched for papers mentioning radiotherapy (or associated terms); artificial intelligence (or associated terms, such as deep learning, CNN, or similar); and outcome, prediction, prognosis or toxicity. Studies were considered if they reported on models for predicting cancer patient outcomes after radiotherapy, utilizing deep learning methodology and including complex dose data as part of the prediction model input, either as full 3D dose distribution or after some limited dimensionality reduction (for example as DVH information). We excluded studies which only considered imaging features or a small number of summary dose metrics for prediction.

Introduction

Radiation oncology has a long history of outcome prediction research, manifest by innumerable tumour control probability (TCP) and normal tissue complication probability (NTCP) modelling studies. The field has been a forerunner in personalized medicine [1]; not just through careful radiotherapy plan adjustment to individual patient anatomy, but also using predictive models to guide a wide range of decision making, including treatment modality allocation [2–4] and individual treatment optimization [5,6]. As such, radiation oncology is well positioned to gain from development and clinical implementation of predictive artificial intelligence (AI) for further personalisation.

Classical dose-response and outcome modelling in radiation oncology has necessarily relied on crude data reduction techniques to represent complex treatment data, especially for anatomy and radiation dose representation. These models have provided real clinical benefit, be it by guiding conformal radiotherapy plan optimisation or guiding clinical trial development, but have generally been phenomenological in nature. The vast majority of studies have reduced radiation dose to simple one-dimensional representations ('metrics') for specific anatomical structures, such as mean absorbed dose, volume receiving a certain dose threshold, etc, based on clinician outlining on the treatment planning CT scan. These dose metrics have been related to outcome through statistical models, typically some form of generalized linear modelling, potentially taking patient characteristics or other clinical factors into account. This approach partly reflects the inability of most traditional statistical modelling methodology to handle highly complex, multidimensional radiotherapy treatment data. Standard approaches not only run afoul of the collinear nature of radiation dose data, but also largely discard the spatial information. This has proven a particular problem for NTCP prediction, especially given the conformal dose distributions seen with modern treatment techniques.

Driven by these issues, machine learning methodologies have been utilised in the attempt to better understand patient and treatment factors and their complex relationships with radiotherapy outcomes [7,8]. Separate to these efforts, methods have been developed to explore spatial dose dependencies, particularly in normal tissue. These latter approaches fundamentally work by transferring dose distributions onto a common reference geometry, and then comparing dose in responders and non-responders on a voxel-by-voxel level with standard statistical models [9]. They thereby utilise the full 3D dose distribution and are often not reliant on manual segmentation. Some published results are promising, potentially identifying previously unknown radiosensitive organ substructures [10]. There are unresolved questions around spatial data normalisation and statistical significance testing, however [11].

Deep learning methodologies may offer a way to explore spatial dose correlations beyond the individual voxel level, while also integrating patient-specific factors and imaging markers. Deep learning is subset of a broader family of machine learning methods which are based on artificial neural networks. These networks are built from multiple connected layers, where only the input layer – for example an image - and the output layer - often a general classification, such as "toxicity yes/no", or a voxel-wise classification, for example for image segmentation - are visible to the user. Deep learning methods learn a task from training examples by optimising weights on the connections between layers. A convolutional neural network (CNN) is a specific type of neural network where a limited number of neurons in each layer are connected to each other through kernels (sliding filters). The most distinctive features of the convolution operation are: (i) that the weights are noticeably fewer than typical fully-connected neural networks, which alleviates the time and memory-related difficulties whilst training, and reduces the risk of overfitting; (ii) that the convolution is a shift-invariant operation which means the features extracted from the input are somewhat more robust to shift and translation. CNNs have proven highly successful for a broad range of image analysis tasks, and are now widely applied to medical imaging, including for organ segmentation, object detection, registration and classification. Classification using CNNs, one of the most researched tasks, has achieved performance rivalling human experts [12]. Examples include Alzheimer's disease classification based on MR images [13] and pneumonia classification using a deep CNN [14].

Treatment response and outcome prediction can be considered a type of classification task, and the use of individual patient imaging to predict outcomes after radiotherapy has been explored in multiple studies (see for example [15–17]). Outcome prediction after radiotherapy using solely imaging or clinical information is conceptually no different than outcome prediction in other medical fields, and does not utilise the patient-specific dose distribution information available in radiation oncology. Radiation dose distributions are large 3D objects, comparable to anatomical imaging in size and complexity. In other words, radiation dose distributions can be considered just another type of imaging data, either as 2D or 3D information. They should hence lend themselves well to deep learning methodologies, and in particular CNNs. Specifically, such an approach might help shed light on poorly understood questions like spatial radiosensitivity in individual organs, or differences in local control probability across treatment targets.

This overview will focus on the specific challenge of radiation dose data in this setting, and will provide a scoping review of the use of deep learning methodology incorporating radiation dose information for outcome prediction. We will highlight current challenges and emerging opportunities, but also discuss established best practices for prognosis and prediction research which should be followed irrespectively of the modelling methodology.

Overview of published studies

Altogether, our initial search identified 627 papers mentioning radiotherapy; artificial intelligence (or associated terms, such as deep learning, CNN, or similar); and outcome, prediction, prognosis or toxicity. Of those, 96 abstracts were screened, and 18 papers selected for full-text review. Ten papers reported on deep learning models for outcome prediction utilising spatial dose information [18–27], see Table 1, while four studies used dimensionally reduced (DVH level) information for prediction [28–31].

Focusing on the former, as summarised in Table 1, the majority used 3D dose matrices directly as input into their CNN architecture to classify patients as experiencing / not experiencing toxicity (i.e. as a binary outcome measure). Zhen et al [18] considered 2D rectum dose surface maps to predict rectal toxicity after cervical cancer treatment, using a methodology which is unlikely to be generalisable to most other organs - the majority of organs do not have shapes which are homeomorphic to a cylinder and thus unfolding to 2D leads to distortions. Ibragimov et al modelled post-SBRT survival and local cancer progression rather than a toxicity endpoint in their 2019 paper [21], and Welch et al considered 3-year locoregional failure for oropharyngeal cancer patients [23]. And finally, Wang et al [25] proposed to predict early voxel-wise FDG-PET response in gross tumour volume (GTV) and clinical target volume (CTV) after radiotherapy for oropharyngeal cancer; the only study not using a single binary per-patient outcome measure. This paper also used 2D axial slices as independent input rather than full 3D images or dose matrices.

Features from CT images, with or without treatment planning contours, were the most common additional data type used for prediction. Men et al [20] demonstrated improved performance for a network utilising dose, CT and structures, compared to individual data types alone, studying xerostomia prediction in 784 head and neck cancer patients; although the addition of clinicians' contours made relatively little difference. Wang et al [25], discussed above, used PET images as input, alongside dose and CT. Liang et al [27] used functional lung ventilation images, derived from 4D-CT, and functional dose distributions (weighting the dose with the ventilation image) as input to their model to predict radiation pneumonitis; with combined dose and ventilation images performing better than models using only individual data types. Finally, several studies looked at the addition of patient-level clinical factors, although with somewhat inconsistent conclusions as to whether the combined deep CNN models outperformed simple models with clinical factors alone [19,21,23].

For categorical output prediction (i.e. all studies but Wang et al), most works used a relatively standard CNN architecture; i.e. started with convolutional layers to extract features, down-sampling layers after convolution layers to reduce the data dimension and one or two fully-connected layers at the end to generate the output. The majority of studies used pre-trained convolutional layers for feature extraction, some with additional fine-tuning on the radiotherapy specific data. A number of different approaches were used to handle multiple input data modalities. For the addition of patient-level clinical factors, studies typically considered these as input to one of the final fully connected layers. Ibragimov et al. [19] proposed a multi-path neural network: one path with the input of dose image and other path with the input of other treatment features, with the ultimate features extracted by each path concatenated and the result passed through two FC layers. More variation was seen in the methods used to handle multimodality 3D input: Ibragimov et al [22] simply concatenated CT and dose to a single input matrix per patient. Men et al [20] proposed to use a 3D residual CNN (rCNN) with three inputs (CT images, 3D dose, structure set). Their network had three convolution blocks which extracted the features of the three inputs separately and the summation of extracted features were subsequently passed through the rest of the network. In contrast, Welch et al [23] and Yan et al [26] used multimodality input as different channels for the same set of convolutional layers. Liang et al [27] explicitly explored whether extracting features separately before combining was better than adding all data as separate channels in a single input matrix. They found a slightly better performance with separate feature extraction.

The second group of studies focused on DVH information rather than full 3D dose distributions for outcome prediction. In these papers, neural networks were generally designed with fully connected layers (not convolutional layers) due to the lack of spatial information and low-dimensional data. The architecture could thus be simple, with only two or three layers [28,29,31] or hybrid (i.e. a combination of networks) [30]. The input was mainly a vector of dose features, with other clinical factors added: Compared to the full 2D & 3D networks discussed above, the simple 1D structure of DVH data allows for straightforward concatenation with other, non-dosimetric, features; for example patient characteristics, treatment details, etc. There are potential advantages of this approach in terms of lower data requirements as well as memory usage for model training. Additionally, it retains some of the main benefits of artificial neural networks compared to other machine learning techniques: Firstly, a neural network with an appropriate architecture can model any well-behaved function, with arbitrary nonlinear dependencies. Secondly, the collinearity problem for input data does not affect neural networks to the same extent as for traditional machine learning, where highly collinear data can result in increased variance of model parameter estimates [32]. Thus, whilst the lack of spatial information for DVH based approaches is an obvious disadvantage, CNN based approaches using 3D data will typically not be able to predict critical dose levels due to internal normalisation; there is thus considerably scope for complimentary studies of this type, which may still offer insights related to absolute dose effects.

Table 1: Studies utilising full (spatial) dose information for outcome prediction with deep learning models

Reference	Patient cohort	Dose information	Other input factors	Summary of model architecture	Predicted outcome	Spatial dose dependence qualification	Train / validation / test	Comments
Zhen 2017 [18]	42 cervical cancer patients	2D surface dose flattened from 3D dose	N/A	Pre-trained 2D CNN: 16 convolutional layers followed by max pooling. 3 FC layers with softmax activation function	Rectal toxicity grade ≥2	2D toxicity risk maps (gradient-weighted class activation maps)	10-fold CV LOOCV	10-fold-CV AUC = 0.70. LOOCV AUC = 0.89. Dose to superior rectum might be particularly associated with rectal toxicity.
Ibragimov 2018 [19]	125 patients treated with liver SBRT	3D dose distributions	Non-dosimetric features used to train a separate FC network	Pre-trained 3D CNN: Three sets of convolution layers with dropouts, two max-pooling layers, and 2 FC layers	Late hepatobiliary toxicities grade ≥3	Individual 3D toxicity risk map (saliency maps, created by systematically varying dose input)	20-fold CV	AUC for CNN alone = 0.79. AUC for combination of ANN and CNN = 0.85. Almost two times fewer false-positive compared to DVH-based methods. Toxicity risk for proximal PV was two times higher than left PV. No correlation between dose delivered to central HBT and hepatobiliary toxicity.
Men 2019 [20]	784 SCC head & neck cancer patients	3D dose distributions	3D CT, parotid structures	Separate initial 3D convolution layer for each input (dose, CT, structure set), followed by four deeper bottleneck architecture 3D layers, pooling, fully connected, and softmax loss layers	Late xerostomia grade ≥2	N/A	80% training, 10% validation, 10% test	AUC = 0.84 for full model (dose, CT, contours); better than logistic regression model with dose & clinical factors. Predictive performance generally better with all three inputs, although contours made relatively small difference.
Ibragimov 2019 [21]	120 patients treated with liver SBRT (same cohort as Ibragimov 2018)	3D dose distributions	Non-dosimetric (clinical) features	Multi-path network: Network from Ibragimov 2018 with input of 3D dose plan, combined with a 3-layers FC network with the input of non-dosimetric features	Post-SBRT survival and local cancer progression	Individual 3D risk map (created by systematically varying dose input and	10-fold CV	3D dose and numerical clinical features combined outperformed separate models. The highest risk for negative outcome was related to the dose received by caudate lobe.

						tumour location)		
Ibragimov 2020 [22]	122 patients treated with liver SBRT (same cohort as Ibragimov 2018)	3D dose distributions	3D CT	Pre-trained 3D CNN: 10 convolutional layers, in which the 8-mid layers are residual, followed by FC layer at the end; input is a concatenation of 3D dose and 3D CT images	Late hepatobiliary toxicities grade ≥3	Individual 3D toxicity risk map (saliency maps, created by systematically varying dose input)	20-fold-CV	AUC = 0.73 for prediction. 3D maps allowed for localisation of high toxicity risk for eight regions of liver, PV and HBT. Highest risk region was HBT.
Welch 2020 [23]	160 oropharyngeal cancer patients	3D dose distributions	3D CT, contours, clinical features	Three-channel 3D CNN with dose, CT and structures: Three convolution layers followed by BN and maxpooling; one FC layer (combining output of CNN with clinical factors) with softmax at the end	Locoregional failure at 3 years	N/A	10-fold CV	CNN trained with dose, CT and clinical features performed worse (precision recall AUC = 0.32) than with clinical features alone (PR-AUC = 0.36), and generally worse than user-driven machine learning models.
Liang 2020 [24]	70 NSCLC patients	3D dose distributions	N/A	3D CNN with 5 convolutional layers (pre-trained trained for multi-frame video classification) followed by maxpooling layers and 2 FC layers at the end	Radiation pneumonitis grade ≥2	Guided gradient-weighted class activation maps to find spatial features characteristic for patients with/without toxicity	50-times random 10-fold CV	CNN trained on dose AUC = 0.842; higher than logistic regression models using dose and/or clinical factors. Low-dose region of contralateral lung and high-dose region of ipsilateral lung were strongly correlated with grade ≥ 2 and grade < 2 radiation pneumonitis cases, respectively.
Wang 2020 [25]	66 oropharyngeal cancer patients	2D dose distributions on axial slices	2D CT, 2D FDG-PET images (both on axial slices)	3D network, taking concatenated 2D PET/CT and dose as input; 8 convolutional layers (with constant image dimensions); loss function prioritising GTV/CTV	2D axial PET images at mid-treatment (20Gy out of 70Gy)	N/A	61 patients for training. 5 patients for testing	Predicted mean SUV CTV/GTV values 3.50/1.41 compared to ground truth values of 3.57/1.51. Average 5%/10 mm 2D gamma test pass rate 92%.

Yang 2021 [26]	52 post-prostatectomy patients	3D dose distributions	3D CT scans	2 channels (dose, CT, both cropped to either bladder or rectum); 3 convolutional layers (pretrained as part of an autoencoder network) followed by maxpooling and FC layer	Acute patient-reported urinary and bowel symptoms; worst score during treatment above/below cut-off value		39 patients for training (31 training, 8 evaluation), with 5-fold CV; 13 patients for testing	No useful model could be found for bladder symptoms (median accuracy 38%). Model for change in rectal symptoms had accuracy 74%.
Liang 2021 [27]	217 thoracic cancer patients	3D dose distributions	Ventilation image from 4D-CT; functional dose distribution (weighting dose with ventilation image)	Pre-trained 3D CNN with 5 layers (see Liang 2020) to extract features from each input dataset, followed by feature filtering and selection	Radiation pneumonitis grade ≥2	N/A	5-fold CV for training/testing; nested sampling used to split training dataset into hyper-tuning/validation.	Combining dose and ventilation information (AUC = 0.87) outperformed models using dose, ventilation, or functional dose alone.

SBRT: Stereotactic body radiotherapy. SCC: Squamous cell carcinoma. NSCLC: Non-small cell lung cancer. FDG-PET: Fluorodeoxyglucose positron emission tomography. GTV: Gross tumour volume. CTV: Clinical target volume. CNN: Convolutional neural network. FC layer: Fully-connected layer, where all the neurons in the layer are connected to all the neurons in the next layer. FC Network: Neural network where all the layers are fully-connected. K-fold CV: K fold cross validation is when data is divided to K fold, (K-1) folds are used for training and one fold is used for testing. This is repeated K time so all data are used in training and testing. The reported evaluation metric is the average for each testing fold. LOOCV: Leave One Out Cross Validation. One set of patient data is kept in reserve for testing, and the model optimised on the remainder. The average of the performance is computed for all possible permutations. PV: Portal vein. HBT: Hepatobiliary tract.

Current challenges and opportunities

Use of deep learning for toxicity and tumour response prediction after radiotherapy is clearly in its infancy, with only just over a dozen published papers in the last 2-3 years. A number of methodological issues are specific to the application of deep learning to radiotherapy and/or toxicity prediction, and it is worth considering these in more detail. Oncology outcome data, and in particular early and late toxicity data, tend to be more complex than the typically binary data used for classification in most 'traditional' CNN architectures. The majority of papers summarised in this review considered only clinician scored toxicity, with a single paper predicting acute patient reported symptoms [26]. All of them dichotomised their toxicity data prior to model development, and this step causes significant loss of information: Toxicity data are nearly always more complex; whether it is ordinal grading to represent increasing severity of side effects, as typically used for clinician toxicity scoring, or the more complex and often continuous scales which have been used to capture patient reported outcomes. Traditional NTCP modelling has also struggled with fully utilising this information, but progress has been made for example in use of ordinal regression models [33]. Several machine learning methods have been modified to handle the ordinal regression problem and for neural networks *ranking learning* models have been proposed; these are typically traditional neural networks with different formulation [34–36]. Generally, these models are based on one of two ideas: Either converting the ordinal problem into pairwise binary classification or changing the network architecture in order to learn multi-thresholds for ordinal classification. This has for example been applied to early diagnosis and classification of Alzheimer's disease [37] and to grading of ulcer severity in patients with Crohn's disease [38]. However, the proposed implementations of the ordinal classification problem can result in lower accuracy of learning when some grades are rare, as will typically be the case for the severe end of a toxicity scale. Further work is needed to explored whether these approaches are appropriate for toxicity data. Additionally, time to events and censoring should optimally be taken into account for late toxicity as well as long-term tumour-related outcomes. A number of relatively recent methodological developments have allowed for modelling of survival outcomes from imaging data based on CNNs combined with Cox regression [39–41] or Fine-Gray regression for competing risks [42]. The paper by Cui et al [31] very elegantly demonstrates how this might be applied to radiation oncology outcomes, with competing toxicity and tumour control endpoints, but we have yet to see a study combining full spatial dose information with actuarial data through deep learning. Similarly, no work so far has looked at prediction of time series data, such as toxicity profiles or tumour regression over time, from radiotherapy dose.

Modern radiotherapy treatment data is inherently multimodal, with combined CT images and dose distributions as the minimum dataset – and with the potential to include additional structural and functional imaging, for example MRI and PET. Although dose can be considered on its own for outcome prediction, the majority of studies reviewed here included both dose distributions and imaging in their models. Several approaches were used, including multi-channel networks and initial separate convolutional layers for each modality [20,22,23,25,26]. With the increasing use of online adaptive radiotherapy, there might also be a need to implement methods for time series analysis, considering multiple dose distributions from adaptive planning simultaneously. These methods may extend deep learning models that classify videos [43] to account for reduced temporal and increased spatial dimensions of CT images and dose. Incorporation of multimodality inputs increase the input size and generally requires access to larger GPU memories and more computational resources. To handle large input sizes, recent deep learning developments apply convolutional layers over smaller patches extracted from the original images (for example histopathological images [44]). These methodological developments could potentially be of interest to radiotherapy outcome prediction, as could other aspects of multimodal machine learning [45,46]. However, with the use of smaller patches, there may be a risk of losing more distant spatial context.

Radiotherapy is somewhat unique in medicine in that it allows for spatial modulation of the active agent (the radiation dose). As such, traditional NTCP and TCP modelling has focused on interpretability of dose-volume effects, to allow for translation into treatment plan optimisation. This may also explain why more complex machine learning approaches have seen limited uptake, and why simple models – such as dose constraints based on mean organ dose – still prevail. Deep learning neural networks do not inherently provide interpretability – weights between connecting layers do not have easily interpretable meaning – but there are numerous methods available to help explain CNN models. Gradient-weighted class activation mapping (Grad-CAM) is a technique for CNN interpretation which highlights input regions that are *'important'* for the prediction [47]. Grad-CAM back propagates the masked gradient of the predicted class with respect to the feature maps that carry the same spatial information of the original image. These gradient maps are then rectified and aggregated to generate the overall class activation map, highlighting image areas that are salient for the predicted class. Zhen et al [18] and Liang et al [24] used Grad-CAM to explain the neural network behaviour for the outcome prediction [20]. However, although they show that the CNN can extract important regions of the input image, it is not clear how regions are related to the output. Furthermore, when analysing spatial dose data, Grad-CAM and feature map analyses do not tell us whether certain regions are more sensitive to change in dose than others; i.e. they cannot be used to explore spatial variations in radiosensitivity. To first order, toxicity will always depend on delivered dose, so the most important regions will be those most often irradiated. The interesting information however, is sensitivity to change in dose, which will be tissue and dose-level dependent. Ibragimov et al [19,21,22] used the concept of gradients of input features to provide interpretability and explore regional dose sensitivity: After training the CNN, they created two new, artificial dose distributions for each pixel *x* in the dose distribution; one with the value of *x* increased and one with the value of *x* decreased. The two new dose plans were separately fed to the CNN and the subtraction of the predicted outputs provided an indication of dose sensitivity for the pixel in question. In other words, the map of pixel-wise output changes provides a map of sensitivity to change in dose for the outcome under investigation. Beyond radiotherapy, the question of interpretable neural networks is an extremely active field of research [48]. It is not yet clear, however, how to go from toxicity maps (or similar measures) to actionable information that can be used directly for treatment plan optimisation, the way that DVH metrics are currently used in clinical practice.

The studies reviewed here used small, retrospective and single institutional datasets; with notable exceptions by Men et al, who used data from 784 patients with head and neck cancer from the RTOG 0522 trial [20], and Cui et al [31], who developed models for radiation pneumonitis on institutional data and externally validated those models on data from 327 patients treated on RTOG 0617. Sufficient and good quality training data is a key requirement for deep learning, and radiotherapy is in no way unique in this aspect - data scarcity is one of the biggest challenges for deep learning development with medical data in general. Three main approaches have been used to alleviate this issue: The first is to use transfer learning, which utilises information from a previously learned task to pre-train a network and improve the performance on the goal task, typically reducing the amount of required training data. Most of the radiotherapy papers discussed in this review used transfer learning. The second is use of data augmentation; a technique which adds slight modifications, for example rotation and scaling, to the existing data to generate new data, as used by Yang et al [26]. Whilst an attractive solution to domain specific data-scarcity, transfer learning relies on the assumption that the 'source-task' is sufficiently similar to the 'target-task', both in terms of input data and predictive output. CNNs operating on image data are often assumed to be transferrable due to the similarity of low-level (detail) features across many types of image. Retraining of later CNN layers or fully connected prediction heads is common for task-adaption. However, radiotherapy dose data are quite unlike other images, lacking sharp features and exhibiting smoother variation over larger receptive fields. Therefore an optimal CNN architecture for radiotherapy dose might require larger filter sizes than a typical image-CNN, reducing the efficacy of transfer learning. Furthermore, the pre-learned low-level image features may be absent from radiotherapy dose data and relevant features will not be well represented in the generic network weights. The combination of these effects leads to a risk of large task dissimilarity, which can lead to 'negative transfer' where transfer learning is detrimental relative to direct learning on the small domain-specific dataset [49]. These issues are likely to be particularly prominent in radiotherapy dose analyses and careful investigation is required as to the suitability of transfer learning in this domain. The third is generating synthesized data, using generative models – including generative neural networks [50] and over sampling techniques such as SMOTE [51] – which are able to generate fake data with the same schema and statistical properties as their "real" counterpart. Due to the higher complexity for 3D medical image-like data, this technique is not very common for medical imaging data augmentation. These three methods can partly counter data availability issues, but are not a complete solution, as the statistical data distribution is fundamentally determined by the underlying real samples. Just as for more traditional NTCP modelling approaches, there is a need to facilitate access to high-quality multi-institutional datasets, for example through national and international repositories and secondary use of trial datasets [52] Notably, none of the studies in Table 1 used external datasets for model validation, and it is thus unclear how well the results will generalise. Curated and diverse multi-institutional datasets could provide an independent and general source of validation data for future deep learning-based models.

Use of machine learning for clinical outcome prediction in oncology generally suffers from poor reporting [53], and this is also reflected in the radiotherapy-specific literature. The vast majority of papers fail to explicitly report according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines [54], despite these being the accepted standard for prognostic model reporting. The much-awaited AI-specific TRIPOD statement will hopefully be published in the near future [55], and will help provide a link between the machine learning and the epidemiology-biostatistics modelling communities. Until then, all items in the original TRIPOD statement are applicable to deep learning models, even when the terminology may at times differ. Other ‘best practices’ from classic outcome prediction modelling should also be strived for, including use of prospectively registered study protocols and data analysis plans [56] and publication of full models and code for independent validation – both of which were also lacking in the reviewed papers. It is worth noting that many of these issues were already highlighted in the QUANTEC papers over a decade ago [52,57].

Conclusion

Deep learning methodologies may offer a better way to model complex dose-response relationships in radiation oncology, while also integrating imaging and clinical features. They may also help shed light on poorly understood questions like spatial variation in radiosensitivity in individual normal tissues or tumours. The published literature on use of deep learning for radiotherapy outcome prediction is relatively scarce, however, and suffers from a number of general methodological issues, including small patient cohorts and lack of external validation. In addition, there are a number of issues specific to the intersection of radiotherapy outcome modelling and deep learning, for example translation of model developments into treatment plan optimisation. These latter issues will likely require efforts from AI experts with radiotherapy domain specific knowledge to solve.

Funding acknowledgements

The authors would like to acknowledge Cancer Research UK funding (grant C19942/A28832) for the Leeds Radiotherapy Research Centre of Excellence (RadNet). Ane Appelt is supported by Yorkshire Cancer Research Academic Fellowship funding (grant L389AA). Alexandra Gilbert is supported by CRUK Clinical trial Fellowship funding (grant C52878/A28301).

References

[1] Baumann M, Krause M, Overgaard J, Debus J, Bentzen SM, Daartz J, et al. Radiation oncology in the era of precision medicine. Nature Reviews Cancer 2016;16:234–49. https://doi.org/10.1038/nrc.2016.18.

[2] 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. Radiotherapy and Oncology 2013;107:267–73. https://doi.org/10.1016/j.radonc.2013.05.007.

[3] Quik EH, Feenstra TL, Postmus D, Slotman BJ, Leemans CR, Krabbe PFM, et al. Individual patient information to select patients for different radiation techniques. European Journal of Cancer 2016;62:18–27. https://doi.org/10.1016/j.ejca.2016.04.008.

[4] Tambas M, Steenbakkers RJHM, van der Laan HP, Wolters AM, Kierkels RGJ, Scandurra D, et al. First experience with model-based selection of head and neck cancer patients for proton therapy. Radiotherapy and Oncology 2020;151:206–13. https://doi.org/10.1016/j.radonc.2020.07.056.

[5] Haslett K, Bayman N, Franks K, Groom N, Harden S v, Harris C, et al. Isotoxic Intensity Modulated Radiation Therapy in Stage III Non-Small Cell Lung Cancer: A Feasibility Study. International Journal of Radiation Oncology, Biology, Physics 2021;109:1341–8. https://doi.org/10.1016/j.ijrobp.2020.11.040.

[6] Vogelius IR, Håkansson K, Due AK, Aznar MC, Berthelsen AK, Kristensen C a, et al. Failure-probability driven dose painting. Medical Physics 2013;40:081717. https://doi.org/10.1118/1.4816308.

[7] Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning applications in radiation oncology. Physics and Imaging in Radiation Oncology 2021;19:13–24. https://doi.org/10.1016/j.phro.2021.05.007.

[8] Luo Y, Chen S, Valdes G. Machine learning for radiation outcome modeling and prediction. Medical Physics 2020;47:e178–84. https://doi.org/10.1002/mp.13570.

[9] Ebert MA, Gulliford S, Acosta O, de Crevoisier R, McNutt T, Heemsbergen WD, et al. Spatial descriptions of radiotherapy dose: normal tissue complication models and statistical associations. Physics in Medicine & Biology 2021;66:12TR01. https://doi.org/10.1088/1361-6560/ac0681.

[10] Green A, Vasquez Osorio E, Aznar MC, McWilliam A, van Herk M. Image Based Data Mining Using Per-voxel Cox Regression. Frontiers in Oncology 2020;10:1178. https://doi.org/10.3389/fonc.2020.01178.

[11] Shortall J, Palma G, Mistry H, Vasquez Osorio E, McWilliam A, Choudhury A, et al. Flogging a Dead Salmon? Reduced Dose Posterior to Prostate Correlates With Increased PSA Progression in Voxel-Based Analysis of 3 Randomized Phase 3 Trials. International Journal of Radiation Oncology, Biology, Physics 2021;110:696–9. https://doi.org/10.1016/j.ijrobp.2021.01.017.

[12] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018;172:1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.

[13] Wang S-H, Phillips P, Sui Y, Liu B, Yang M, Cheng H. Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling. Journal of Medical Systems 2018;42:85. https://doi.org/10.1007/s10916-018-0932-7.

[14] Yadav SS, Jadhav SM. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data 2019;6:113. https://doi.org/10.1186/s40537-019-0276-2.

[15] Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. Scientific Reports 2019;9:2764. https://doi.org/10.1038/s41598-019-39206-1.

[16] Jin C, Yu H, Ke J, Ding P, Yi Y, Jiang X, et al. Predicting treatment response from longitudinal images using multi-task deep learning. Nature Communications 2021;12:1851. https://doi.org/10.1038/s41467-021-22188-y.

[17] Bibault J-E, Giraud P, Housset M, Durdux C, Taieb J, Berger A, et al. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Scientific Reports 2018;8:12611. https://doi.org/10.1038/s41598-018-30657-6.

[18] 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. Physics in Medicine and Biology 2017;62:8246–63. https://doi.org/10.1088/1361-6560/aa8d09.

[19] Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A, Xing L. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. Medical Physics 2018;45:4763–74. https://doi.org/10.1002/mp.13122.

[20] Men K, Geng H, Zhong H, Fan Y, Lin A, Xiao Y. A Deep Learning Model for Predicting Xerostomia Due to Radiation Therapy for Head and Neck Squamous Cell Carcinoma in the RTOG 0522 Clinical Trial. International Journal of Radiation Oncology, Biology, Physics 2019;105:440–7. https://doi.org/10.1016/j.ijrobp.2019.06.009.

[21] Ibragimov B, Toesca DAS, Yuan Y, Koong AC, Chang DT, Xing L. Neural Networks for Deep Radiotherapy Dose Analysis and Prediction of Liver SBRT Outcomes. IEEE Journal of Biomedical and Health Informatics 2019;23:1821–33. https://doi.org/10.1109/JBHI.2019.2904078.

[22] Ibragimov B, Toesca DAS, Chang DT, Yuan Y, Koong AC, Xing L, et al. Deep learning for identification of critical regions associated with toxicities after liver stereotactic body radiation therapy. Medical Physics 2020;47. https://doi.org/10.1002/mp.14235.

[23] Welch ML, McIntosh C, McNiven A, Huang SH, Zhang B-B, Wee L, et al. User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. Physica Medica 2020;70:145–52. https://doi.org/10.1016/j.ejmp.2020.01.027.

[24] Liang B, Tian Y, Chen X, Yan H, Yan L, Zhang T, et al. Prediction of Radiation Pneumonitis With Dose Distribution: A Convolutional Neural Network (CNN) Based Model. Frontiers in Oncology 2019;9:1500. https://doi.org/10.3389/fonc.2019.01500.

[25] Wang C, Liu C, Chang Y, Lafata K, Cui Y, Zhang J, et al. Dose-Distribution-Driven PET Image-Based Outcome Prediction (DDD-PIOP): A Deep Learning Study for Oropharyngeal Cancer IMRT Application. Frontiers in Oncology 2020;10:1592. https://doi.org/10.3389/fonc.2020.01592.

[26] Yang Z, Olszewski D, He C, Pintea G, Lian J, Chou T, et al. Machine learning and statistical prediction of patient quality-of-life after prostate radiation therapy. Computers in Biology and Medicine 2021;129:104127. https://doi.org/10.1016/j.compbiomed.2020.104127.

[27] Bin L, Yuan T, Zhaohui S, Wenting R, Zhiqiang L, Peng H, et al. A deep learning-based dual-omics prediction model for radiation pneumonitis. Medical Physics 2021;48:6247–56. https://doi.org/10.1002/mp.15079.

[28] Pella A, Cambria R, Riboldi M, Jereczek-Fossa BA, Fodor C, Zerini D, et al. Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy. Medical Physics 2011;38:2859–67. https://doi.org/10.1118/1.3582947.

[29] Chan ST, Ruan D, Shaverdian N, Raghavan G, Cao M, Lee P. Effect of Radiation Doses to the Heart on Survival for Stereotactic Ablative Radiotherapy for Early-stage Non-Small-cell Lung Cancer: An Artificial Neural Network Approach. Clinical Lung Cancer 2020;21:136-144.e1. https://doi.org/10.1016/j.cllc.2019.10.010.

[30] Zhu C, Lin SH, Jiang X, Xiang Y, Belal Z, Jun G, et al. A novel deep learning model using dosimetric and clinical information for grade 4 radiotherapy-induced lymphopenia prediction. Physics in Medicine and Biology 2020;65:035014. https://doi.org/10.1088/1361-6560/ab63b6.

[31] Cui S, ten Haken RK, el Naqa I. Integrating Multiomics Information in Deep Learning Architectures for Joint Actuarial Outcome Prediction in Non-Small Cell Lung Cancer Patients After Radiation Therapy. International Journal of Radiation Oncology, Biology, Physics 2021;110:893–904. https://doi.org/10.1016/j.ijrobp.2021.01.042.

[32] de Veaux Richard D. and Ungar LH. Multicollinearity: A tale of two nonparametric regressions. In: Cheeseman P. and Oldford RW, editor. Selecting Models from Data, New York, NY: Springer New York; 1994, p. 393–402.

[33] Appelt AL, Bentzen SM, Jakobsen A, Vogelius IR. Dose-response of acute urinary toxicity of long-course preoperative chemoradiotherapy for rectal cancer. Acta Oncologica (Stockholm, Sweden) 2015;54:179–86. https://doi.org/10.3109/0284186X.2014.923933.

[34] da Costa Joaquim Pinto and Cardoso JS. Classification of Ordinal Data Using Neural Networks. In: Gama João and Camacho R and BPB and JAM and TL, editor. Machine Learning: ECML 2005, Berlin, Heidelberg: Springer Berlin Heidelberg; 2005, p. 690–7.

[35] Fernandez-Navarro F, Riccardi A, Carloni S. Ordinal Neural Networks Without Iterative Tuning. IEEE Transactions on Neural Networks and Learning Systems 2014;25:2075–85. https://doi.org/10.1109/TNNLS.2014.2304976.

[36] Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters 2020;140:325–31. https://doi.org/10.1016/j.patrec.2020.11.008.

[37] Li H, Habes M, Fan Y. Deep Ordinal Ranking for Multi-Category Diagnosis of Alzheimer's Disease using Hippocampal MRI data 2017. http://arxiv.org/abs/1709.01599

[38] Barash Y, Azaria L, Soffer S, Margalit Yehuda R, Shlomi O, Ben-Horin S, et al. Ulcer severity grading in video capsule images of patients with Crohn's disease: an ordinal neural network solution. Gastrointestinal Endoscopy 2021;93:187–92. https://doi.org/10.1016/j.gie.2020.05.066.

[39] Li H, Boimel P, Janopaul-Naylor J, Zhong H, Xiao Y, Ben-Josef E, et al. DEEP CONVOLUTIONAL NEURAL NETWORKS FOR IMAGING DATA BASED SURVIVAL ANALYSIS OF RECTAL CANCER. Proceedings IEEE International Symposium on Biomedical Imaging 2019;2019:846–9. https://doi.org/10.1109/ISBI.2019.8759301.

[40] Hao L, Kim J, Kwon S, Ha I do. Deep Learning-Based Survival Analysis for High-Dimensional Survival Data. Mathematics 2021;9:1244. https://doi.org/10.3390/math9111244.

[41] Shao W, Wang T, Huang Z, Han Z, Zhang J, Huang K. Weakly Supervised Deep Ordinal Cox Model for Survival Prediction from Whole-slide Pathological Images. IEEE Transactions on Medical Imaging 2021;PP. https://doi.org/10.1109/TMI.2021.3097319.

[42] Lee C, Zame W, Yoon J, van der Schaar M. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proceedings of the AAAI Conference on Artificial Intelligence; 32(1), 2018.

[43] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-Scale Video Classification with Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE; 2014, p. 1725–32. https://doi.org/10.1109/CVPR.2014.223.

[44] Sudharshan PJ, Petitjean C, Spanhol F, Oliveira LE, Heutte L, Honeine P. Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications 2019;117:103–11. https://doi.org/10.1016/j.eswa.2018.09.049.

[45] Baltrusaitis T, Ahuja C, Morency L-P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 2019;41:423–43. https://doi.org/10.1109/TPAMI.2018.2798607.

[46] Bayoudh K, Knani R, Hamdaoui F, Mtibaa A. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer 2021. https://doi.org/10.1007/s00371-021-02166-7.

[47] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision 2020;128:336–59. https://doi.org/10.1007/s11263-019-01228-7.

[48] Zhang Q, Zhu S. Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering 2018;19:27–39. https://doi.org/10.1631/FITEE.1700808.

[49] Wang Z, Dai Z, Póczos B, Carbonell J. Characterizing and Avoiding Negative Transfer. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, p. 11285–94. https://doi.org/10.1109/CVPR.2019.01155.

[50] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, Cambridge, MA, USA: MIT Press; 2014, p. 2672–80.

[51] Chawla N v., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002;16:321–57. https://doi.org/10.1613/jair.953.

[52] Deasy JO, Bentzen SM, Jackson A, ten Haken RK, Yorke ED, Constine LS, et al. Improving normal tissue complication probability models: the need to adopt a “data-pooling” culture. International Journal of Radiation Oncology, Biology, Physics 2010;76:S151-4. https://doi.org/10.1016/j.ijrobp.2009.06.094.

[53] Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. Journal of Clinical Epidemiology 2021;138:60–72. https://doi.org/10.1016/j.jclinepi.2021.06.024.

[54] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Medicine 2015;13:1–10. https://doi.org/10.1186/s12916-014-0241-z.

[55] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11:e048008. https://doi.org/10.1136/bmjopen-2020-048008.

[56] Thor M, Oh JH, Apte AP, Deasy JO. Registering Study Analysis Plans (SAPs) Before Dissecting Your Data-Updating and Standardizing Outcome Modeling. Frontiers in Oncology 2020;10:978. https://doi.org/10.3389/fonc.2020.00978.

[57] Jackson A, Marks LB, Bentzen SM, Eisbruch A, Yorke ED, ten Haken RK, et al. The lessons of QUANTEC: recommendations for reporting and gathering data on dose-volume dependencies of treatment outcome. International Journal of Radiation Oncology, Biology, Physics 2010;76:S155-60. https://doi.org/10.1016/j.ijrobp.2009.08.074.