Physics and Imaging in Radiation Oncology 16 (2020) 149–155

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

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

Original Research Article

Automatic evaluation of contours in radiotherapy planning utilising conformity indices and machine learning

Samsara Terparia [a,*], Romaana Mir [b], Yat Tsang [a,b], Catharine H Clark [b,c,d], Rushil Patel [b]

[a] Radiotherapy Physics, Mount Vernon Cancer Centre, Northwood, UK
[b] NIHR Radiotherapy Trials Quality Assurance Group, Mount Vernon Cancer Centre, Northwood, UK
[c] Radiotherapy Physics, University College London Hospital, London, UK
[d] National Physical Laboratory, Teddington, UK

ARTICLE INFO

Keywords:
Machine learning
Conformity index
Quality assurance
Interobserver variation
Delineation
SABR

ABSTRACT

Background and purpose: Peer-review of Target Volume (TV) and Organ at Risk (OAR) contours in radiotherapy planning are typically conducted visually; this can be time consuming and subject to interobserver variation. This study investigated automatic evaluation of contouring using conformity indices and supervised machine learning.

Methods: A total of 393 contours from 253 Stereotactic Ablative Body Radiotherapy (SABR) benchmark cases (adrenal gland, liver, pelvic lymph node and spine), delineated by 132 clinicians from 25 centres, were visually evaluated for conformity against gold standard contours. Contours were scored as "pass" or "fail" on visual peer review and six Conformity Indices (CIs) were applied. CI values were mapped to pass/fail scores for each contour and used to train supervised machine learning models. A 5-fold cross validation method was employed to determine the predictive accuracies of each model.

Results: The stomach structure produced models with the highest predictive accuracy overall (96% using Support Vector Machine and Ensemble models), whilst the liver GTV produced models with the lowest predictive accuracy (76% using Logistic Regression). Predictive accuracies across all models ranged from 68–96% (68–87% for TV and 71–96% for OARs).

Conclusions: Although a final visual review by an experienced clinician is still required, the automatic contour evaluation method could reduce the time for benchmark case reviews by identifying gross contouring errors. This method could be successfully implemented to support departmental training and the continuous assessment of outlining for clinical staff in the peer-review process, to reduce interobserver variability in contouring and improve interpretation of radiological anatomy.

1. Introduction

Interobserver variability amongst clinicians in outlining of Target Volumes (TV) and Organs at Risk (OARs) is a challenge in radiotherapy. Inaccuracies in TV and OAR contour delineation may impact on both tumour control and normal tissue toxicities [1]. Though The Royal College of Radiologists recommends that all radiotherapy departments should have processes that enable optimal TV delineation and peer-review [2], no formal outlining training exists for clinical staff at present.

The assessment of the contours can be conducted both visually and quantitatively using Conformity Indices (CIs) against pass/fail criteria [3]. Visual inspection is currently the most common approach however, this is time consuming as it requires a meticulous slice-by-slice evaluation of TV and OARs. There is no current standard in radiotherapy Quality Assurance (QA) to quantitatively evaluate the agreement of contours. CIs are mathematical metrics which can be used to quantify contour conformity with a Gold Standard (GS). Some of these include the Dice Similarity Coefficient (DSC), Jaccard Conformity Index (JCI), van't Riet Index (VRI), Geographical Miss Index (GMI), Discordance Index (DI) and Hausdorff Distance (HD) [4,5]. Whilst these CIs provide a quantitative evaluation of contour conformity, without standardised assessment criteria this data is limited in use.

Manually establishing pass/fail criteria for a wide range of CI data

* Corresponding author.
E-mail address: sami.terparia@nhs.net (S. Terparia).

https://doi.org/10.1016/j.phro.2020.10.008
Received 29 February 2020; Received in revised form 20 October 2020; Accepted 21 October 2020
Available online 1 December 2020

Table 1
Table of benchmarks, structures and imaging provided to clinicians to aid delineation in the pre-trial outlining benchmark exercises.

SABR Benchmark	Spine	Pelvic Lymph Node	Adrenal Gland	Liver
Number of Cases	71	85	44	53
Imaging Modalities Provided to aid Delineation	CT, MRI (T1 and T2 weighted)	Contrast enhanced CT, MRI (T2 weighted)	Contrast enhanced CT	4DCT, 3DCT, MRI (T2 weighted)
Structures Considered	GTV	GTV	Liver Stomach	GTV Liver Stomach Oesophagus Heart

across multiple TV and OAR structures can be laborious. Machine Learning (ML) models may be employed to decipher statistical patterns between a qualitative pass/fail score and quantitative DSC, JCI, VRI, HD, GMI and DI values for each investigator contour.

Supervised ML is branch of Artificial Intelligence (AI) where computer algorithms learn from prior experience [6]. Supervised ML algorithms use training data with known input (predictors) and output (responses) values, to detect patterns and correlation through the learning process [7], which can then be used to predict whether investigator contours "pass" or "fail" pre-trial outlining exercises. Whilst several studies have investigated the use of AI for auto-segmentation contouring in radiotherapy planning [8–10], the use of ML to assess TV and OAR contour conformity is limited.

This study examined the feasibility of using qualitative pass/fail criteria and CIs to develop ML models to assess the compliance of TV and OAR contour delineations according to a standardised protocol.

2. Methods

2.1. Data pre-processing

The United Kingdom National Institute for Health Research (NIHR) Radiotherapy Trials Quality Assurance (RTTQA) Group implements pre-trial outlining benchmark exercises for clinical trials to reduce interobserver variability, ensuring that patients are treated according to protocol guidelines. Before a new clinician (an investigator) can recruit patients to a clinical trial, the investigator must delineate TV and OAR contours for each benchmark and submit the contours to the RTTQA Group for review. The contours are reviewed to determine outlining compliance against GS contours, which have been delineated by the Chief Investigator, or agreed as a consensus by the Trial Management Group (TMG) [11].

A total of 393 investigator contours from 253 Stereotactic Ablative Body Radiotherapy (SABR) pre-trial outlining benchmark exercises, delineated by 132 clinicians from 25 centres across the UK, that had previously been reviewed by the RTTQA Group SABR credentialing programme [12,13] were utilised for this study. Outlining benchmark cases were created using data from previously treated SABR patients who had consented for their data to be used for education and training purposes. The RTTQA Group provided investigators with detailed contouring atlases over a range of imaging modalities to aid contour delineation (Table 1).

Each investigator contour was imported into Velocity (v4.01 Varian Medical Systems). An experienced RTTQA clinician was blinded to the investigator submitting the contours and visually compared each investigator contour in transverse, coronal, and sagittal planes to the GS, evaluating if the respective contours encompassed the TV or OAR structure in full and if the OAR contour extended into an adjacent OAR. Contours were scored as either "pass" or a "fail" dependending on agreement to the GS; an example of a "fail" included an OAR structure not contoured to completion (e.g. oesophageal wall not contoured in entirety) or an OAR structure extending into an adjacent structure (e.g. submitted stomach contour includes the GS duodenum contour). This scoring was required in order to label the investigator contours into the "pass" or "fail" categories, as required by the training process in supervised ML.

The investigator and GS contours for each benchmark were then exported from Velocity and imported into CERR (Computational Environment for Radiotherapy Research) to calculate JCI, DSC, VRI, HD, GMI and DI values; formulae for the CIs used are detailed in Fig. S1. CERR is an open source application written in a Matlab environment, for viewing and analysing radiotherapy data. CI values for each investigator contour were then mapped to their corresponding pass/fail score.

2.2. Supervised machine learning

CI data and mapped pass/fail scores for each investigator contour were imported into The Classification Learner Application in Matlab to apply supervised ML using the classification technique, which involves building and evaluating ML models for discrete responses that can be classified into categories [14]. The JCI, DSC, VRI, HD, GMI and DI values were used as predictors and the corresponding pass/fail scores were used as responses. The aim was to allow each ML model to decipher trends between CIs and pass/fail scores, such that they could then predict a pass/fail outcome based on the CI values of unseen investigator contours.

When training ML models, it is recommended to have at least five to ten samples per predictor [15]. The requirement on the minimum number of samples per predictor can vary based on the complexity of the data and model. Due to the simplicity of our proposed models and the limitation of the small sample size, five samples per predictor was chosen in this study and this required a minimum of thirty investigator contours per model. Datasets are typically split 70%/30% into training and testing datasets [16], where the training dataset is used to train the model and the testing dataset is used to validate the model to determine its predictive accuracy, sensitivity and specificity. For smaller datasets, a k-fold cross validation technique can be applied to test and validate the data and produce each model's predictive accuracy, sensitivity and specificity. To maximise the training dataset size, a 5-fold cross validation technique for testing and validation was used, where $k = 5$ was chosen as this has shown to yield test error rate estimates with low bias [17].

Suitability of the ML model is highly dependent on the nature of the dataset. In addition to this, the weighting of each predictor is determined by each specific model. To mitigate this effect, the dataset should be applied to more than one model to establish which is the most appropriate [18]. The models considered were Decision Tree, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Ensemble (summarised in Table S1), as these have been used in previous radiation oncology related studies [19–21].

All CI predictors, except the HD, were dimensionless quantities that took a value between zero and one; the HD was given as a distance value in centimetres. The KNN model uses a distance metric to calculate the closeness of data. When training the KNN model, the predictor values should be scaled to be of similar magnitude to allow all predictors to contribute equally [14]. Therefore, a normalised HD that took a value between zero and one, was used instead of the HD to train KNN models to produce more reliable results.

Models were trained using CI values mapped to pass/fail scores in three approaches: (i) for all 393 contours as a single group; (ii) for TV structures and OAR structures as two separate groups; (iii) for individual structure type.

Non-parametric Kruskal-Wallis tests were used to investigate stastically significant differences in the variations of (i) all CI values across each structure type; (ii) ML model predictive accuracies, sensitivities

Table 2
Table showing resulting 5-fold predictive accuracies, sensitivities and specificities of trained models by structure and algorithm type.

Structure Type (number of "Pass" contours/total number of contours)		Machine Learning Model accuracy (%) (Sensitivity %/Specificity %)				
		Tree	Logistic Regression	Support Vector Machine	K- Nearest Neighbour	Ensemble
All (242/393–62%)		77 (75/79)	72 (93/39)	80 (88/66)	78 (89/58)	78 (81/74)
TV (148/209–71%)		84 (88/71)	80 (93/49)	80 (94/48)	80 (89/56)	79 (88/61)
All OAR (94/184–51%)		78 (77/76)	71 (87/54)	78 (84/76)	79 (87/71)	82 (85/79)
TV	Liver GTV (34/53–64%)	68 (79/55)	76 (82/65)	70 (97/25)	68 (82/45)	68 (88/35)
	Node GTV (66/85–78%)	81 (91/47)	81 (94/37)	85 (94/42)	86 (97/37)	82 (94/42)
	Spine GTV (48/71–68%)	79 (90/70)	83 (94/57)	86 (94/70)	87 (92/78)	83 (85/78)
OAR	Liver (41/68–60%)	78 (88/63)	78 (85/67)	84 (85/78)	85 (90/78)	81 (81/82)
	Stomach (25/67–37%)	93 (88/93)	92 (92/93)	96 (100/93)	91 (96/88)	96 (96/95)
	Oesophagus (20/32–63%)	81 (85/75)	72 (85/50)	75 (90/50)	84 (75/92)	84 (90/75)
	Heart (8/17–47%)	71 (88/56)	88 (88/89)	94 (100/89)	88 (88/89)	82 (75/89)

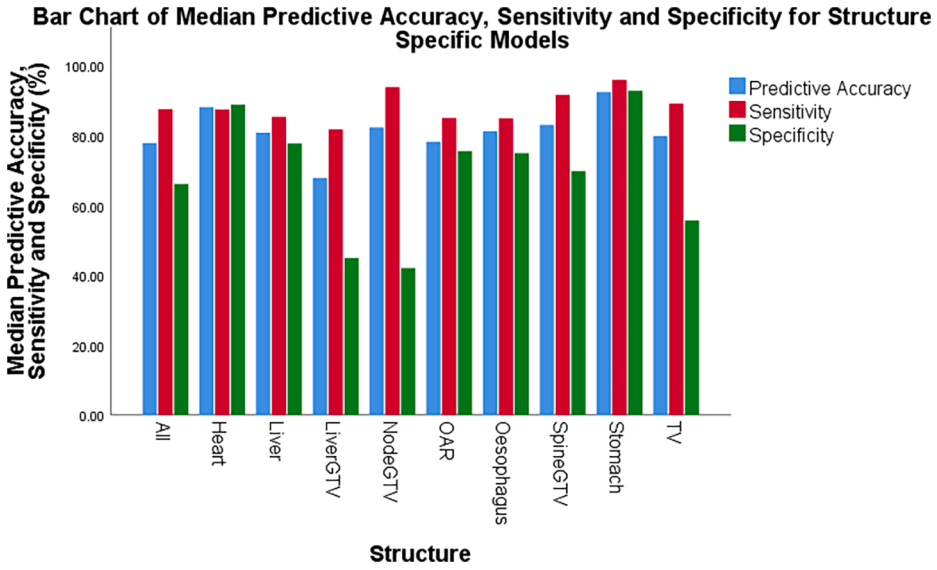

Fig. 1. Bar chart demonstrating the median predictive accuracies, sensitivities and specificities obtained across all trained ML models.

and specificities among the groups of all contours, TV and OARs; (iii) ML model predictive accuracies, sensitivities and specificities among the groups of liver GTV, node GTV, spine GTV, liver, stomach, oesophagus and heart.

3. Results

Resulting predictive accuracies, sensitivities and specificities of trained models are shown in Table 2. Predictive accuracies observed across all models in this study ranged from 68–96%, where models for TV and OARs ranged from 68–87% and 71–96%, respectively. No statistically significant differences were found in the variations of predictive accuracies ($p = 0.08$), sensitivities ($p = 0.11$), nor specificities ($p = 0.50$) amongst the models created for all contours, TV and OARs. For liver GTV, node GTV, spine GTV, liver, stomach, oesophagus and heart models, there were statistically significant differences in the variations in predictive accuracies, sensitivities and specificities ($p < 0.05$). Fig. 1 shows that by training structure specific models, higher predictive accuracies, sensitives and specificities may be achieved. There were statistically significant differences in variations of JCI, DSC, VRI, HD, GMI and DI values across each structure type ($p < 0.05$). The relationship between each CI is illustrated in Figs. S2–S6.

The stomach structure produced models with the highest predictive accuracy (96% using SVM and Ensemble models) whilst the liver GTV produced models with the lowest predictive accuracy (highest accuracy of 76% using Logistic Regression). The Receiver Operator Characteristic (ROC) curves (Fig. 2) demonstrate the sensitivity (100%) and specificity (93%) for the stomach SVM model, with an Area Under Curve (AUC) of 0.97, and sensitivity (82%) and specificity (65%) for the liver GTV Logistic Regression model with an AUC of 0.84. These two models are fitted in Fig. 3 and demonstrate that the data is closer together for the liver GTV Logistic Regression model in comparison to the stomach SVM model and that the liver GTV Logistic Regression model incorrectly classified more contours than the stomach SVM model. The liver GTV structure had the highest rate of re-submissions and in some instances, re-submissions were still deemed a "fail". Qualitative comparisons of stomach and liver GTV investigator contour submissions to their GS are demonstrated in Fig. 4.

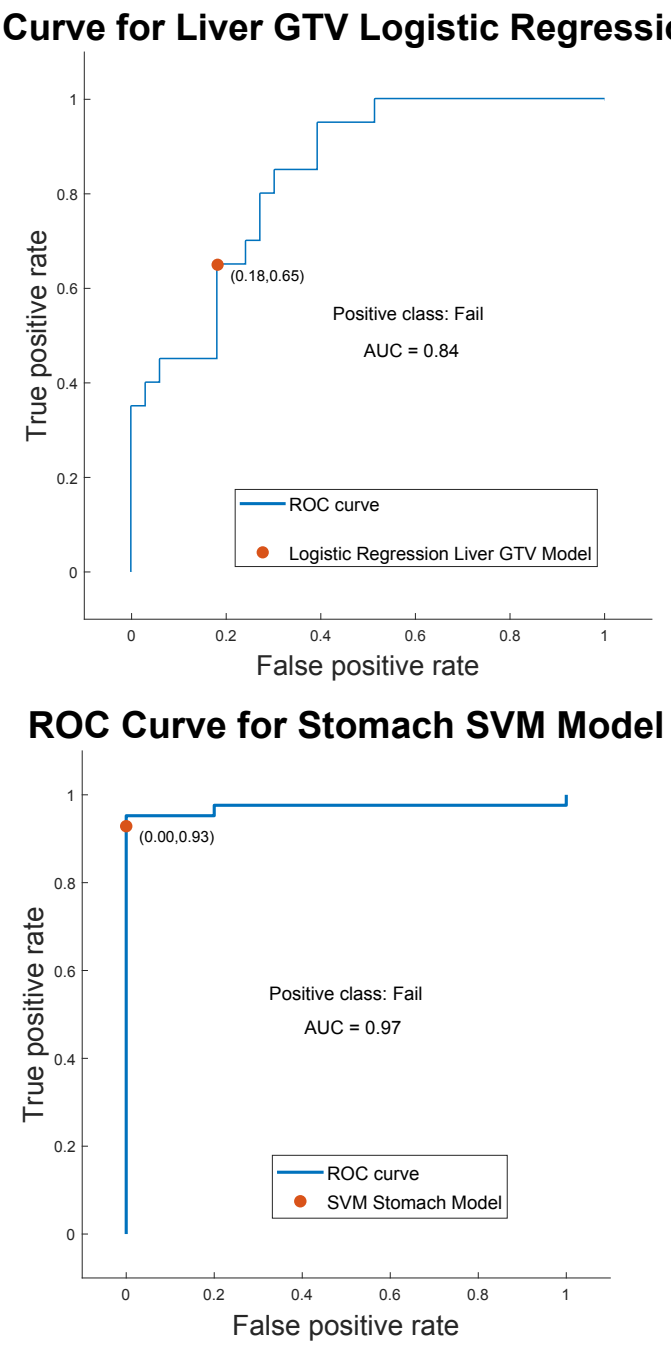

Fig. 2. Receiver Operator Characteristic (ROC) curves for Liver GTV Logistic Regression Model (Top) and Stomach SVM Model (Bottom). The Area Under Curve (AUC) represents the model's overall ability to correctly classify structures into each category.

4. Discussion

This study investigated the feasibility of building an assessment tool, based on conformity indices and supervised ML, to evaluate contouring. Our study found that this is achievable for selected structures and ML models. No statistically significant differences in model predictive accuracy, sensitivity nor specificity amongst the groups of all contours, all TV and all OARs, were identified. By dividing the dataset into structure type, statistically significant differences in predictive accuracy, sensitivity and specificity are seen. Furthermore, statistically significant differences were found in CI values amongst different structures and may explain why higher specificities, and in some cases sensitivities, were observed when models were trained by structure type. This may have aided the ML models in correctly classifying data in to "pass" and "fail" categories, in comparison to the dataset as a whole.

The stomach structures provided trained models with the highest predictive accuracies, sensitivities and specificities. This may be due to several clinicians incorrectly including the first superior slices of the duodenum as the inferior part of the stomach, which would have had an effect on the CI values in terms of over-contouring. Inclusion of the duodenum for stomach contours were automatically scored as a fail. As this was a common issue, the ML models learnt to associate the CI values of these over-contoured stomach structures with their corresponding "fail" score during the training process, validated during the 5-fold cross validation process. This is likely to be the reason for the high specificities attained for the stomach models, as shown in the ROC curve (Fig. 2).

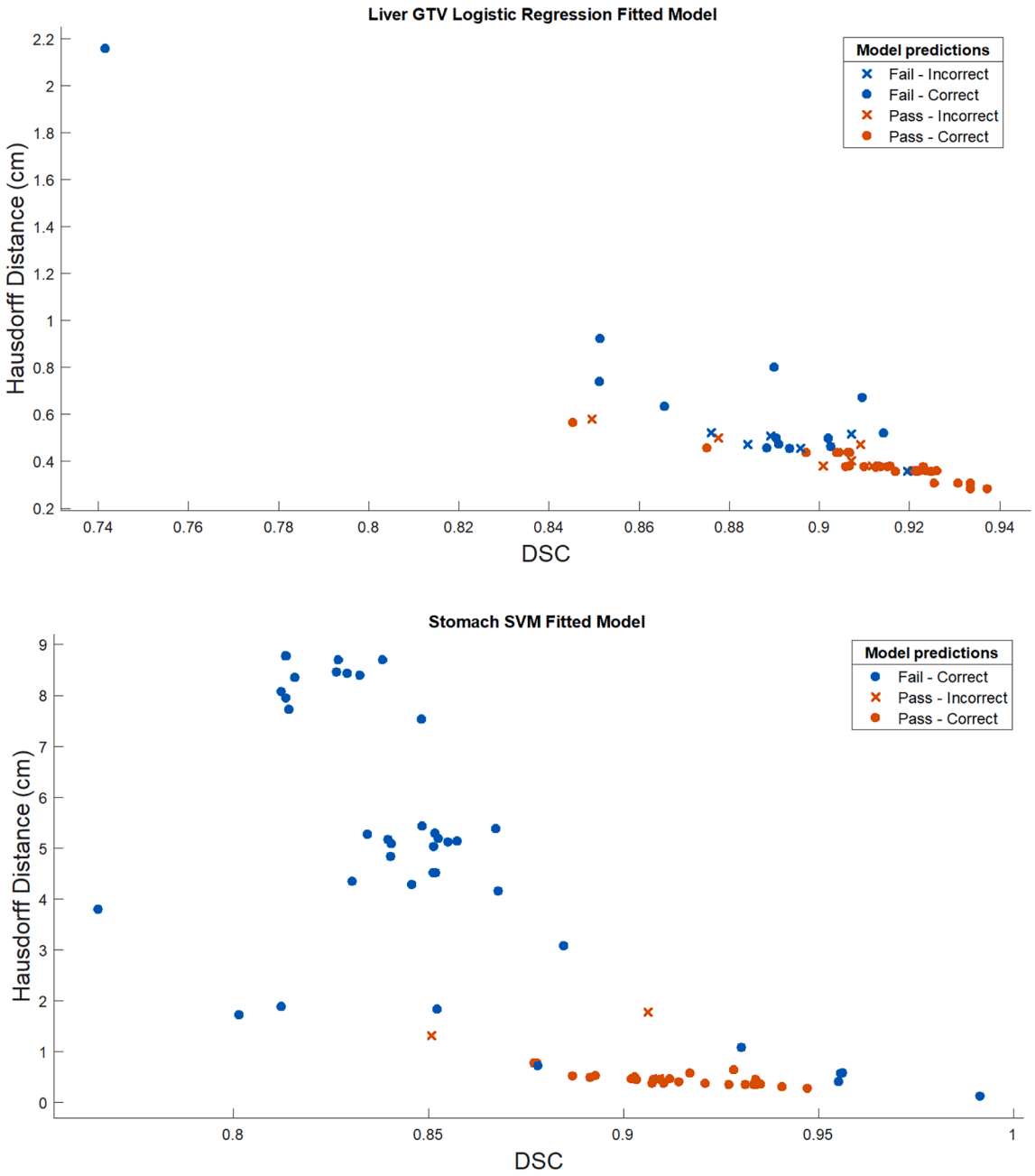

Fig. 3. Scatter plots showing Liver GTV Logistic Regression Model (Top) and Stomach SVM Model (Bottom), fitted with DSC against Hausdorff distance. Blue crosses indicate where the model has predicted an investigator contour as a "fail", but was visually scored as a "pass". Blue dots indicate where the model has correctly predicted an investigator contour as a "fail". Red crosses indicate where the model has predicted an investigator contour as a "pass", but was visually scored as a "fail". Red dots indicate where the model has correctly predicted an investigator contour as a "pass".

Whilst the heart SVM model achieved a predictive accuracy of 94%, this was based on seventeen investigator contours which violates the five sample per predictor rule. One study [6] created models that violated this rule and produced adequate results, however, this heart SVM model may be limited in use as the high predicative accuracy could have arisen as a result overfitting. This is where a ML model learns and represents the training dataset very well, but cannot accurately deal with unseen data.

The liver GTV structure produced models with the lowest predictive accuracies and specificities. Fig. 3 demonstrates that the liver GTV investigator contours were fairly well conformed to the GS and perhaps, the closeness of the data may have resulted in the liver GTV models struggling to correctly classify "pass" and "fail" contours. This is indicative that more examples may be needed for this model to improve its ability to categorise "pass" and "fail" structures.

Using CIs alone to assess conformity is limited, as CI values vary between structures. Currently, no criteria exist to define structure specific CI values that demonstrate high and low conformity. In this study, we mapped CI values to a "pass" or "fail" score and used ML to allow models to determine these criteria for themselves, for future investigator contours to be assessed against. Careful consideration was taken when selecting this dataset as the limitations of ML originate from data quality, where models are subject to "garbage in, garbage out". Investigator contours had previously been reviewed by the RTTQA Group and underwent a second evaluation by an experienced RTTQA clinician to be scored. Determining the optimal distribution of data in each "pass" and "fail" category to train ML models is complex, as there are arguments that favour both a balanced [22] and unbalanced [23] split. Using

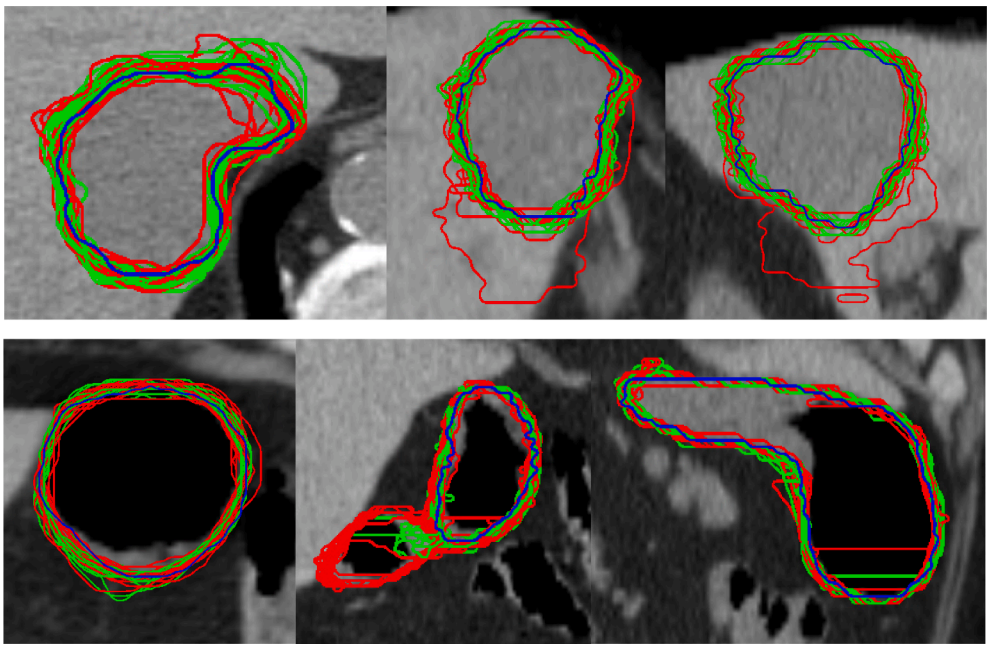

Fig. 4. Comparison of passed (green) and failed (red) Liver GTV (top) and Stomach contours (bottom) in transverse, coronal and sagittal views (left to right). The Gold Standard is outlined in blue.

different distributions of data for a given ML model have shown to affect the learning process [24] which may have an effect on predictive accuracies.

The predictive accuracies, sensitivities and specificities produced in this study were based on a 5-fold cross validation technique, used in favour of a training and testing dataset split, in order to increase the training dataset size to produce maximally robust models. As a result, the true sensitivity and specificity of each model could not be established, rendering it unclear as to how well these models would perform on unseen data. The more representative examples of "pass" and "fail" data, the more robust and accurate a trained model will be. More data is required to refine and test these models before they can be used routinely. Alternatively, if small datasets are unavoidable, fewer predictors could be used for training. For example, as the JCI, DSC and VRI are mathematically similar (Figs. S2–S6), two of these could be omitted as predictors. With four predictors, the minimum number of samples required is reduced to twenty however, having too few predictors leads to underfitting. This is a phenomenon in ML where too few predictors result in oversimplified models that cannot identify data trends [16]. Models presented here are relatively simple with few predictors and removing predictors may increase the risk of underfitting. Whilst some CIs used here are mathematically correlated, using them together does not impact upon a model's ability to make predictions of new observations on unseen data [25] and helps increase model complexity to reduce the likelihood of underfitting. Alternatively, the use of other types of mathematical metrics, such as the surface DSC [26] and added path length [27], could be investigated to determine their suitability for this purpose.

Whilst this method cannot yet replace a clinical review, these models may be used to pre-evaluate investigator contours prior to clinical review, to highlight gross contouring errors to the reviewer. The balance of sensitivities and specificities should be considered when choosing a ML model for this purpose. Models with specificities close to 100% are ideal as this results in a low false positive rate, reducing the number of investigator contours being scored as a "pass" by the model that would have failed a visual review. This expedites the review process whilst minimising the number of false positive contours being accepted in the credentialing process. The RTTQA Group SABR Expansion Programme expect a further 300 outlining benchmark submissions that will enable testing of these models on unseen data to determine their true sensitivities and specificities. A larger data set would enable investigation of alternative ML models and the effect of varying the number of predictors and distributions of data in each category on predictive accuracy, sensitivity and specificity. This will allow fine-tuning of the models to handle unseen data more effectively, improving their reliability so that they can be used to replace aspects of the clinical review.

The automatic contour evaulation method could be successfully implemented in supporting departmental training and the continuous assessment of outlining for clinical staff. The collection of completed outlining exercises, that have been visually evaluated against a GS, can be mapped to their corresponding CI values and used to create and train departmental ML models. These models may then be used to assess outlining performed by staff in training. CI values calculated for the investigator contour and GS would be provided to the models to identify failed contours that require further visual review, expediting the peer-review process.

At present no formal outlining training exists for clinical staff. The Royal College of Radiologists recommends that all radiotherapy departments should have processes that enable optimal TV delineation and peer-review. Developing a QA programme such as this could provide a proactive and standardised approach to comply with these recommendations to support departmental training and the continuous assessment of outlining for clinical staff in the peer-review process, reducing interobserver variability in contouring and improving interpretation of radiological anatomy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The RTTQA Group is funded by the National Institute for Health Research. The authors would like to thank all of the centres and clinical oncologists that undertook the SABR pre-trial outlining benchmark exercises.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2020.10.008.

References

[1] Simões R, Wortel G, Wiersma TG, Janssen TM, van der Heide UA, Remeijer P. Geometrical and dosimetric evaluation of breast target volume auto-contouring. Phys Imag Radiat Oncol 2019;12:38–43. https://doi.org/10.1016/j.phro.2019.11.003.

[2] The Royal College of Radiologists. Radiotherapy target volume definition and peer review-RCR guidance, https://www.rcr.ac.uk/system/files/publication/field_publication_files/bfco172_peer_review_outlining.pdf; 2017 [accessed August 5, 2020].

[3] Gwynne S, Gilson D, Dickson J, McAleer S, Radhakrishna G. Evaluating target volume delineation in the era of precision radiotherapy: FRCR revalidation and beyond. Clin Oncol 2017;29:436–8. https://doi.org/10.1016/j.clon.2017.01.045.

[4] Gwynne S, Spezi E, Wills L, Nixon L, Hurt C, Joseph G, et al. Toward semi-automated assessment of target volume delineation in radiotherapy trials: the SCOPE 1 pretrial test case. Int J Radiat Oncol Biol Phys 2012;84:1037–42. https://doi.org/10.1016/j.ijrobp.2012.01.094.

[5] Li X, Zhang YY, Shi YH, Zhou LH, Zhen X. Evaluation of deformable image registration for contour propagation between CT and cone-beam CT images in adaptive head and neck radiotherapy. Technol Health Care 2016;24:S747–55. https://doi.org/10.3233/THC-161204.

[6] Panesar SS, D'Souza RN, Yeh FC, Fernandez-Miranda JC. Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. World Neurosurg X 2019;2:100012. https://doi.org/10.1016/j.wnsx.2019.100012.

[7] Boon IS, Au Yong TPT, Boon CS. Assessing the role of artificial intelligence (AI) in clinical oncology: utility of machine learning in radiotherapy target volume delineation. Medicines 2018;5:131. https://doi.org/10.3390/medicines5040131.

[8] Brunenberg EJL, Steinseifer IK, van den Bosch S, Kaanders JHAM, Brouwer CL, Gooding MJ, et al. External validation of deep learning-based contouring of head and neck organs at risk. Phys Imag Radiat Oncol 2020;15:8–15. https://doi.org/10.1016/j.phro.2020.06.006.

[9] Gurney-Champion OJ, Kieselmann JP, Wong KH, Ng-Cheng-Hin B, Harrington K, Oelfke U. A convolutional neural network for contouring metastatic lymph nodes on diffusion-weighted magnetic resonance images for assessment of radiotherapy response. Phys Imag Radiat Oncol 2020;15:1–7. https://doi.org/10.1016/j.phro.2020.06.002.

[10] Haq R, Hotca A, Apte A, Rimner A, Deasy JO, Thor M. Cardio-pulmonary substructure segmentation of radiotherapy computed tomography images using convolutional neural networks for clinical outcomes analysis. Phys Imag Radiat Oncol 2020;14:61–6. https://doi.org/10.1016/j.phro.2020.05.009.

[11] Gwynne S, Spezi E, Sebag-Montefiore D, Mukherjee S, Miles E, Conibear J, et al. Improving radiotherapy quality assurance in clinical trials: assessment of target volume delineation of the pre-accrual benchmark case. Br J Radiol 2013;86:20120398. https://doi.org/10.1259/bjr.20120398.

[12] NHS England. Commissioning through evaluation: Standards for the provision of stereotactic ablative radiotherapy, http://www.swscn.org.uk/wp/wp-content/uploads/2014/11/SABR-CtE-Service-Specification-2nd-Sept-2015-final.pdf; 2015 [accessed August 5, 2020].

[13] UK SABR Consortium. Stereotactic Ablative Body Radiation Therapy (SABR): A Resource V6.1.0., https://www.sabr.org.uk/wp-content/uploads/2019/04/SABRconsortium-guidelines-2019-v6.1.0.pdf; 2019 [accessed July 15, 2020].

[14] Kuhn M, Johnson K. Applied predictive modeling. 1st ed. New York: Springer Science & Business Media; 2013.

[15] Somorjai R, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 2003;19:1484–91. https://doi.org/10.1093/bioinformatics/btg182.

[16] Moore MM, Slonimsky E, Long AD, Sze RW, Iyer RS. Machine learning concepts, concerns and opportunities for a pediatric radiologist. Pediatr Radiol 2019;49:509–16. https://doi.org/10.1007/s00247-018-4277-7.

[17] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. 1st ed. New York: Springer Science & Business Media; 2013.

[18] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2006;2:59–77. https://doi.org/10.1177/117693510600200030.

[19] Socarrás Fernández JA, Mönnich D, Leibfarth S, Welz S, Zwanenburg A, Leger S, et al. Comparison of patient stratification by computed tomography radiomics and hypoxia positron emission tomography in head-and-neck cancer radiotherapy. Phys Imag Radiat Oncol 2020;15:52–9. https://doi.org/10.1016/j.phro.2020.07.003.

[20] Mohebian MR, Marateb HR, Mansourian M, Mañanas MA, Mokarian F. A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. Comput Struct Biotechnol J 2017;15:75–85. https://doi.org/10.1016/j.csbj.2016.11.004.

[21] Fernandes CD, Dinh CV, Walraven I, Heijmink SW, Smolic M, van Griethuysen JJM, et al. Biochemical recurrence prediction after radiotherapy for prostate cancer with T2w magnetic resonance imaging radiomic features. Phys Imag Radiat Oncol 2018;7:9–15. https://doi.org/10.1016/j.phro.2018.06.005.

[22] Albisua I, Arbelaitz O, Gurrutxaga I, Lasarguren A, Muguerza J, Pérez JM. The quest for the optimal class distribution: An approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. Prog Artif Intell 2013;2:45–63. https://doi.org/10.1007/s13748-012-0034-6.

[23] Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 2004;6:20–9. https://doi.org/10.1145/1007730.1007735.

[24] Weiss GM, Provost F. The effect of class distribution on classifier learning: an empirical study. Tech Rep 2001. https://doi.org/10.7282/T3-VPFW-SF95.

[25] Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models. 5th ed. New York: McGraw-Hill, Irwin; 2005.

[26] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. ArXiv 2018:1809.04430.

[27] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imag Radiat Oncol 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.