A method to combine target volume data from 3D and 4D planned thoracic radiotherapy patient cohorts for machine learning applications

Radiotherapy and Oncology 126 (2018) 355–361

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Machine-learning in thoracic radiotherapy

A method to combine target volume data from 3D and 4D planned thoracic radiotherapy patient cohorts for machine learning applications

Corinne Johnson [a,b,*], Gareth Price [a,b], Jonathan Khalifa [b,c], Corinne Faivre-Finn [a,b], Andre Dekker [d], Christopher Moore [a,b], Marcel van Herk [a,b]

[a] Manchester Cancer Research Centre, Division of Molecular and Clinical Cancer Science, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Sciences Centre; [b] The Christie NHS Foundation Trust, Manchester Academic Health Sciences Centre, UK; [c] Department of Radiation Oncology, Institut Universitaire du Cancer de Toulouse – Oncopole, France; [d] The MAASTRO Clinic, Maastricht University Medical Centre+, The Netherlands

ARTICLE INFO

Article history:
Received 22 September 2017
Received in revised form 14 November 2017
Accepted 19 November 2017
Available online 6 December 2017

Keywords:
Radiotherapy
Machine-learning
GTV
Lung cancer

ABSTRACT

Background and purpose: The gross tumour volume (GTV) is predictive of clinical outcome and consequently features in many machine-learned models. 4D-planning, however, has prompted substitution of the GTV with the internal gross target volume (iGTV). We present and validate a method to synthesise GTV data from the iGTV, allowing the combination of 3D and 4D planned patient cohorts for modelling.

Material and methods: Expert delineations in 40 non-small cell lung cancer patients were used to develop linear fit and erosion methods to synthesise the GTV volume and shape. Quality was assessed using Dice Similarity Coefficients (DSC) and closest point measurements; by calculating dosimetric features; and by assessing the quality of random forest models built on patient populations with and without synthetic GTVs.

Results: Volume estimates were within the magnitudes of inter-observer delineation variability. Shape comparisons produced mean DSCs of 0.8817 and 0.8584 for upper and lower lobe cases, respectively. A model trained on combined true and synthetic data performed significantly better than models trained on GTV alone, or combined GTV and iGTV data.

Conclusions: Accurate synthesis of GTV size from the iGTV permits the combination of lung cancer patient cohorts, facilitating machine learning applications in thoracic radiotherapy.

Machine learning is increasingly being used in a wide range of fields due to its ability to learn from the surrounding environment and identify relationships and correlations between parameters [1]. It is particularly useful in 'big data' environments where data volumes are too great for humans to perform comprehensive analyses. Within radiotherapy, because of the volume of data generated for each patient, its complexity, and the rise in the number of treatment options, there has been increased interest in using machine-learned prediction models to guide treatment pathway decisions. A number of models applied to lung cancer radiotherapy have been reported in the literature, including the prediction of 2-year survival of Non-Small Cell Lung Cancer (NSCLC) patients [2], dysphagia [3] and radiation pneumonitis [4]. Predictive model performance is essentially dictated by the quality and volume of data upon which they are trained. Generally, model quality improves as the training data more closely represent the population data in which it will be applied, and the larger the dataset the more likely it is to correctly sample that population. As data are often missing in many variables, the more features that are selected for modelling, the smaller the complete set of data becomes. Routine clinical data are often so incomplete that cohort sizes can become very limited for any model requiring more than a handful of features.

Using only these limited complete datasets increases the risk of over-fitting and models may become unrepresentative of the population from which the data are sampled. Cohort sizes can be increased by data sharing initiatives such as the Computer Aided Theragnostics (CAT) project which are building towards distributed learning [5–7]. However, even with such strategies, unless we can be certain that missing data are randomly distributed, the risk of biasing models remains if only complete patient datasets are selected. To avoid the issues associated with the use of complete datasets, typically the approach to missing data is to impute it from the available non-missing data. However, even modern multiple imputation techniques risk biasing models if not used carefully [8], and approximation accuracy decreases as the fraction of missing data increases. It is clear that more attention needs to be

* Corresponding author at: The Christie NHS Foundation Trust, Wilmslow Road, Manchester M20 4BX, UK.
E-mail address: corinne.johnson@physics.cr.man.ac.uk (C. Johnson).

https://doi.org/10.1016/j.radonc.2017.11.015
0167-8140/

paid to prospective data recording, but such efforts even if enacted immediately cannot help improve the quality of current clinical data archives. An alternative is to attempt to synthesise missing data from other sources that are available to us. In this paper we show a robustly validated example of synthesising Gross Tumour Volume (GTV) volume and shape data from the recorded motion-adapted GTV, otherwise known as Internal Gross Target Volume (iGTV), among patients who received radiotherapy for NSCLC.

The GTV, as defined by the International Commission on Radiation Units and Measurements (ICRU) [9], has long been associated with patient outcome due to its close link with tumour stage and the volume of irradiated normal tissue. Recently reported radiotherapy outcome prediction models have confirmed this knowledge by identifying the GTV as a highly predictive feature [10,11]. In addition to the GTV volume itself, factors calculated from its geometry, whether dosimetric measures [12], or, more recently, radiomics features [13,14], have been shown to have predictive potential.

The GTV is historically defined by a clinician contouring on a 3D-CT scan of the patient. However, since the advent of 4D-CT and its uptake in dynamic radiotherapy planning, the GTV it typically replaced by the Internal Target Volume (ITV), defined by the ICRU as the volume encompassing the CTV and its motion [15]. This is dogmatically acquired by contouring the GTV on each 4D-CT phase image, expanding them into individual CTVs, and taking their union as the tumour ITV. Practically, however, the process is often streamlined by contouring the 'motion-adapted GTV', or iGTV, directly on the Maximum Intensity Projection (MIP) of the 4D image set before expanding this to account for microscopic spread [16], removing the need for the GTV to be defined directly.

Lung cancer patients treated at our institution before 2011 were planned using 3D-CT scans. The majority treated after this date will have had a 4D planning CT. There are therefore two distinct lung cancer patient cohorts for whom the recorded measures of tumour size and shape, and all factors that depend upon this, are incompatible. The ability to combine both cohorts would significantly alleviate the missing data problem and enable their direct comparison for e.g. time-ordered validation data partitioning. In this study we hypothesise that it is possible to calculate a synthetic GTV (sGTV) and its associated dosimetric features from the iGTV. Our aim is to develop a concise method to obtain a sGTV, robustly validate it, and demonstrate the utility of this method in improving the accuracy of clinical outcome predictive modelling.

Materials and methods

Patient cohort

We selected a cohort of 40 4D-planned NSCLC patients from routine IMRT data approved for this study by the UK Health Research Authority and The Christie Caldicott information governance committee (ethical approval Ref. 17/NW/0060). These patients were selected to be representative of the tumour location seen in the wider population, by including approximately 1/3 lower lobe and 2/3 upper lobe tumours. 4D-CT was acquired with the 'Philips Bellows' pressure belt system to obtain the surrogate respiratory signal and was reconstructed at 512×512 resolution with 3 mm slice separation and pixel size of 0.97 mm or 1.17 mm. The (ICRU defined) GTV was contoured by an experienced radiation oncologist on the end exhale phase image in addition to the pre-existing MIP iGTV segmentation. The clinician was not blinded to the iGTV.

The patients were randomly split into development ($n = 25$) and validation cohorts ($n = 15$). As lower lobe tumours typically exhibit greater motion due to their proximity to the diaphragm [17,18], upper and lower lobe cases were separated before final cohort assignment.

A further population of 747 NSCLC patients treated at The Christie, for whom either a GTV ($N = 318$) *or* an iGTV ($N = 429$) were contoured, were used to demonstrate the use of the synthesised GTV method for combining data cohorts (the 'modelling cohort'). A cohort of patients treated at the MAASTRO Clinic (Maastricht, The Netherlands), who had their GTVs defined on a 3D CT with the aid of FDG-PET, was used as a blind external validation set ($N = 274$) [2].

Linear fit: volume prediction

Using the development cohort, linear relationships between the iGTV and the GTV volumes were determined separately for upper and lower lobe tumours. The intercept was constrained to be zero, to require a GTV to exist in order for an iGTV to exist. All iGTV and GTV volumes were calculated within the Pinnacle planning system. Synthetic GTV volumes were calculated using these fits for each case within the validation set.

Erosion: volume and contour geometry prediction

Calculation of dosimetric features relies upon the defined GTV shape. The recovery of sGTV geometry from iGTV contours is therefore required for the calculation of such parameters. We use binary morphology operators with kernels estimated directly through comparison of corresponding development cohort iGTVs and GTVs. The differences (in mm) in contour spacing in the horizontal (left–right) and vertical (anterior–posterior) projections from the structures' centres were determined for each slice of the paired contour sets. Additionally, the average slice difference between the two structures was determined. The integer values of the average differences in the horizontal and vertical directions, along with the average slice difference were used to construct 3D elliptical kernels, which were then applied to the validation cohort to generate a sGTV from the iGTV.

Volume comparison

Individualised error thresholds for each validation patient's true GTV volume, which the synthetic GTV should fall within, were estimated by uniformly eroding and dilating the true GTVs by 1 and 2 pixels (approximately 1 mm and 2 mm). No erosion/dilation was performed in the cranial–caudal direction owing to the 3 mm slice thickness, which is larger than the change to be studied. The error tolerance thresholds were chosen as 'acceptable' (2 mm) and 'strict' (1 mm) limits on the synthetic GTV volume, respectively, and were designed to be in accordance with the ICRU recommended uncertainty limits for radiotherapy dose and positional accuracy of 2% and 2 mm [19].

Additionally we calculated the ratios of sGTV/GTV and iGTV/GTV, to compare both the sGTV and the iGTV volumes to the true GTV volume.

Contour comparison

The geometric accuracy of sGTVs in comparison to their true GTVs was assessed using: (a) the Dice Similarity Coefficient (DSC), and (b) determination of the closest point distances between every vertex in corresponding structure contours. The closest point search was performed in both directions (i.e. sGTV → GTV and vice versa) to ensure that crossing contours could not artificially mask the distance to any protrusions. For reference, these metrics were also calculated to compare the iGTV to the true GTV.

Dosimetric comparison

Dosimetric features were calculated for the sGTV and iGTV structures using the clinically validated dose distribution, and the residual difference to the values calculated for the true GTV was determined. Specifically we compared the minimum, mean and maximum dose, the standard deviation of the dose and the dose delivered to 95% of the structure volume (D95). Similar to the volume comparison, strict and acceptable error tolerances were set at 1% and 2% respectively, as per the ICRU recommendations [19]. The difference between the residual doses for the sGTV and the iGTV were compared using paired t-tests.

Random forest model

The efficacy of the synthesised GTVs in combining statically and dynamically planned lung cancer patient populations was explored using the much larger modelling patient cohort. Using the Caret package in R [20], we built a series of Random Forest models to predict 2 year survival of NSCLC patients. Univariate analysis was performed on the sub-set of 3D-planned patients (i.e. the 318 patients with a true GTV defined) for all available variables: age, gender, World Health Organisation Performance Status, T stage, N stage, GTV or sGTV volume, Overall Treatment Time and the GTV or sGTV equivalent dose at 2 Gy (EQD2). All apart from N stage were found to be predictive, and thus were included in each of the models.

The following tests were then performed:

- A model built using just the sGTV Christie patients ($N = 429$) was tested against a validation cohort of true GTV Christie patients ($N = 318$), and vice versa.
- The following set of models were trained and then tested against the unseen external validation cohort ($N = 274$).
  - o A model built without any GTV information ($N = 747$).
  - o A model trained on the true GTV patients alone ($N = 318$).
  - o A model trained on the whole modelling cohort ($N = 747$) but with the GTV values imputed from the existing true GTV data rather than using the synthesised GTV values ($N = 747$).
  - o A model trained on the combined GTV and iGTV data ($N = 747$).
  - o A model trained on the combined GTV and sGTV data ($N = 747$).

All sGTV values were computed using the erosion method. The imputed GTV values were calculated using both simple mean imputation and via a more sophisticated random forest proximity method (using the RandomForestSRC package in R [21–23]). In order to limit the effects of imputation to the GTV only, the cohorts were composed only of patients with complete datasets. Parameter tuning was performed using a gridded search optimised on model accuracy assessed using 5-fold cross-validation in the training data [20]. Model performance in validation data was measured in terms of the Area Under the ROC Curve (AUC), where a value of 0.5 indicates a performance no better than random and a value of 1 indicates perfect classification, together with the model sensitivity/specificity optimised to maximise the prediction Kappa value. Twenty repeats of model training/testing were performed to determine AUC uncertainty and model comparisons were performed using the Welch two-sample t-tests on the resultant sets of AUCs for each model.

Results

Linear fit prediction of volume

The linear fit of the upper lobe patients within the development cohort had a gradient of 1.0833, with constrained intercept 0 (R^2 = 0.9878, RMSE = 6.926), while the linear fit of the lower lobe cases had a gradient of 1.1664, constrained intercept 0 (R^2 = 0.9701, RMSE = 10.073), as shown in Fig. 1.

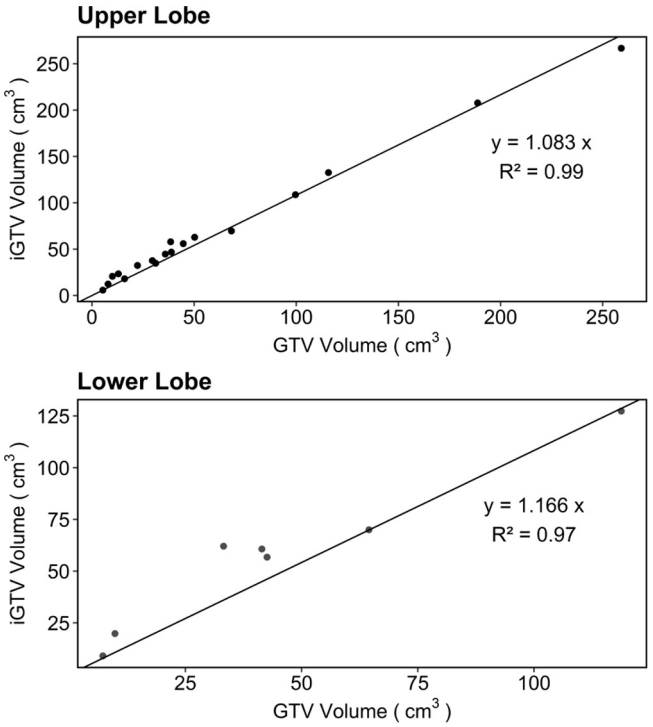

Fig. 1. Linear fits of the true GTV versus the iGTV for both upper (top) and lower (bottom) lobe patients, with the linear equation and the r-squared value of the fit shown on each plot.

Morphological prediction of volume

The average left–right and anterior–posterior distances between the iGTV and GTV contours were 1.81 by 1.91 mm respectively (standard deviation 1.31 and 1.84) for the upper lobe patients, and 2.51 by 3.29 mm (standard deviation 1.48 and 2.49, respectively) for the lower lobe patients of the development cohort. The average slice difference was 1.17 and 1.71 mm respectively (standard deviation 1.59 and 1.35) for upper and lower lobe patients.

The sGTV volumes obtained from the 3D erosion using these kernels are shown in Table 1. The corresponding sGTV/GTV and iGTV/GTV values for the linear fit sGTV, the erosion sGTV and the iGTV are shown in Table 2, where it can be seen that the best volume estimates are achieved by the erosion method, followed by the linear fit. The volumes from the erosion method were therefore used for the random forest models.

Morphological prediction of geometry

The results of the closest point assessment based upon the 3D erosion along with the DSCs are shown in Table 2. The sGTV and iGTV closest point results are comparable, with the mean and range values being statistically similar (mean: $p = 0.38$ and 0.20; range: $p = 0.08$ and 0.98, for upper and lower lobe cases, respectively). Similarly, there is no significant difference between the DSC for the sGTV and for the iGTV for either the upper or lower lobe cases ($p = 0.27$ and 0.10, respectively).

Dosimetric comparison

The mean difference in dose parameters between the sGTV and GTV, and iGTV and GTV, calculated over all validation patients are

Table 1
GTV, iGTV and sGTV volumes.

Patient	True GTV volume (cm^3)	Lower 1 mm confidence interval	Upper 1 mm confidence interval	Lower 2 mm confidence interval	Upper 2 mm confidence interval	Linear Fit sGTV volume (cm^3)	Erosion method sGTV volume (cm^3)	iGTV volume (cm^3)
Upper lobe								
1	73.83	65.75	82.20	56.28	93.24	76.07	73.74	82.41
2	127.70	113.41	142.51	96.58	162.08	137.18	133.10	148.60
3	63.98	58.07	70.05	50.35	78.60	70.06	69.30	75.89
4	3.61	2.41	4.94	1.17	6.76	5.25	4.21	5.69
5	12.64	9.97	15.51	7.22	19.26	13.77	12.10	14.91
6	6.97	5.24	8.84	3.55	11.27	9.97	5.54	10.80
7	19.17	14.17	24.64	8.98	32.30	22.95	19.31	24.86
8	19.55	16.03	23.31	12.07	28.53	25.40	23.37	27.52
Lower lobe								
1	38.06	32.64	43.75	26.72	51.03	46.29	36.94	53.99
2	22.61	19.29	26.10	15.43	30.85	38.24	29.75	44.61
3	54.81	45.69	64.56	35.58	78.01	63.57	54.73	74.15
4	58.35	51.96	65.00	44.09	74.19	66.22	46.44	64.41
5	30.51	23.66	37.84	16.10	48.10	20.61	28.37	47.37
6	85.79	75.50	96.51	63.96	110.24	83.80	65.29	97.75
7	46.50	41.03	52.20	34.61	59.79	51.97	41.48	60.62

True GTV, iGTV and sGTV volumes for each of the upper and lower lobe patients, along with the approximate upper and lower 1 mm and 2 mm confidence intervals. sGTV volumes are given for both the linear fit and the erosion methods.

Table 2
Volume and shape comparison.

Patient	sGTV (Linear Fit)	sGTV (Erosion)					iGTV				
	sGTV/GTV	Mean Difference (mm)	Range (mm)	Standard Deviation of Difference (mm)	Dice Similarity Coefficient	sGTV/GTV	Mean Difference (mm)	Range (mm)	Standard Deviation of Difference (mm)	Dice Similarity Coefficient	iGTV/GTV
Upper lobe											
1	1.03	0.91	0–11.53	0.90	0.94	1.00	0.74	0–12.45	1.17	0.94	1.12
2	1.07	1.40	0–7.79	1.03	0.93	1.04	1.71	0–8.73	1.39	0.92	1.16
3	1.10	1.29	0–8.73	1.13	0.92	1.08	1.20	0–8.18	1.46	0.91	1.19
4	1.45	1.08	0–5.27	0.97	0.82	1.16	1.39	0–4.77	1.26	0.78	1.57
5	1.09	0.91	0–3.27	0.57	0.91	0.96	0.82	0–4.77	0.98	0.92	1.18
6	1.43	1.01	0–5.77	0.99	0.84	1.23	1.46	0–6.08	1.15	0.78	1.54
7	1.19	1.17	0–6.00	0.83	0.85	1.01	1.05	0–6.62	1.30	0.87	1.30
8	1.30	1.75	0–10.16	1.67	0.83	1.20	1.79	0–11.13	2.19	0.83	1.41
Mean	**1.21**	**1.19**	**0–7.32**	**1.01**	**0.88**	**1.08**	**1.27**	**0–7.84**	**1.36**	**0.87**	**1.31**
S.D.	**0.17**	**0.29**	**0–2.75**	**0.31**	**0.05**	**0.10**	**0.38**	**0–2.84**	**0.37**	**0.07**	**0.18**
Lower lobe											
1	1.22	1.55	0–6.31	1.10	0.85	0.97	1.95	0–9.66	1.80	0.83	1.42
2	1.69	3.86	0–21.99	4.69	0.65	1.32	4.13	0–24.88	6.23	0.67	1.97
3	1.16	2.31	0–15.80	2.35	0.81	1.00	1.66	0–14.54	2.43	0.85	1.35
4	0.95	2.04	0–14.43	1.40	0.84	0.80	0.90	0–8.81	1.50	0.95	1.10
5	1.33	2.31	0–20.63	2.24	0.72	0.93	2.03	0–25.33	3.42	0.78	1.55
6	0.98	1.99	0–11.83	1.41	0.83	0.76	0.99	0–6.45	1.42	0.93	1.14
7	1.12	1.69	0–6.29	1.11	0.89	0.89	1.75	0–7.29	1.60	0.86	1.30
Mean	**1.21**	**2.25**	**0–13.90**	**2.04**	**0.80**	**0.95**	**1.92**	**0–13.85**	**2.63**	**0.84**	**1.41**
S.D.	**0.25**	**0.77**	**0–6.25**	**1.27**	**0.08**	**0.18**	**1.07**	**0–8.11**	**1.74**	**0.09**	**0.29**

Volume ratio results are given for the sGTV determined from the linear fits and the erosion methods, as well as the iGTV, for all upper and lower lobe cases. Each is a comparison to the true GTV volume. Additionally, for the sGTV and iGTV, the Dice similarity coefficient and the closest point assessment results are presented, again as compared to the true GTV structures. Closest point assessment results are based upon the difference between the true GTV contours and those predicted from the erosion of the iGTV (sGTV), and additionally for the iGTV itself. Results are given in terms of the mean, range and standard deviation of the closest point distances, determined over the whole structure.

shown in Fig. 2. Generally the dosimetric parameters are similar for the sGTV and iGTV, likely because the dose distribution is designed to be uniform within the PTV. Significant differences were determined for the minimum dose, the maximum dose and the standard deviation (p-values of 0.005, 0.01 and 0.01, respectively).

Random forest model

The model trained on the sGTV only cohort and tested against the true GTV only cohort had an AUC value of 0.71 ± 0.003. The inverse model – trained on the true GTV only cohort and tested against the sGTV only cohort – had an AUC value of 0.73 ± 0.009.

The evaluation parameters of the models tested blindly against the external dataset are shown in Table 3. Using the Welch t-test, all model combinations were found to be statistically different (all with $p < 0.01$) apart from: the true GTV only and the mean imputation cohorts; and the GTV plus iGTV and the random forest imputation cohorts. The best result was found using the sGTV.

Discussion

In this study, we have developed a simple method for calculating a synthetic GTV from the iGTV, with the aim of increasing the

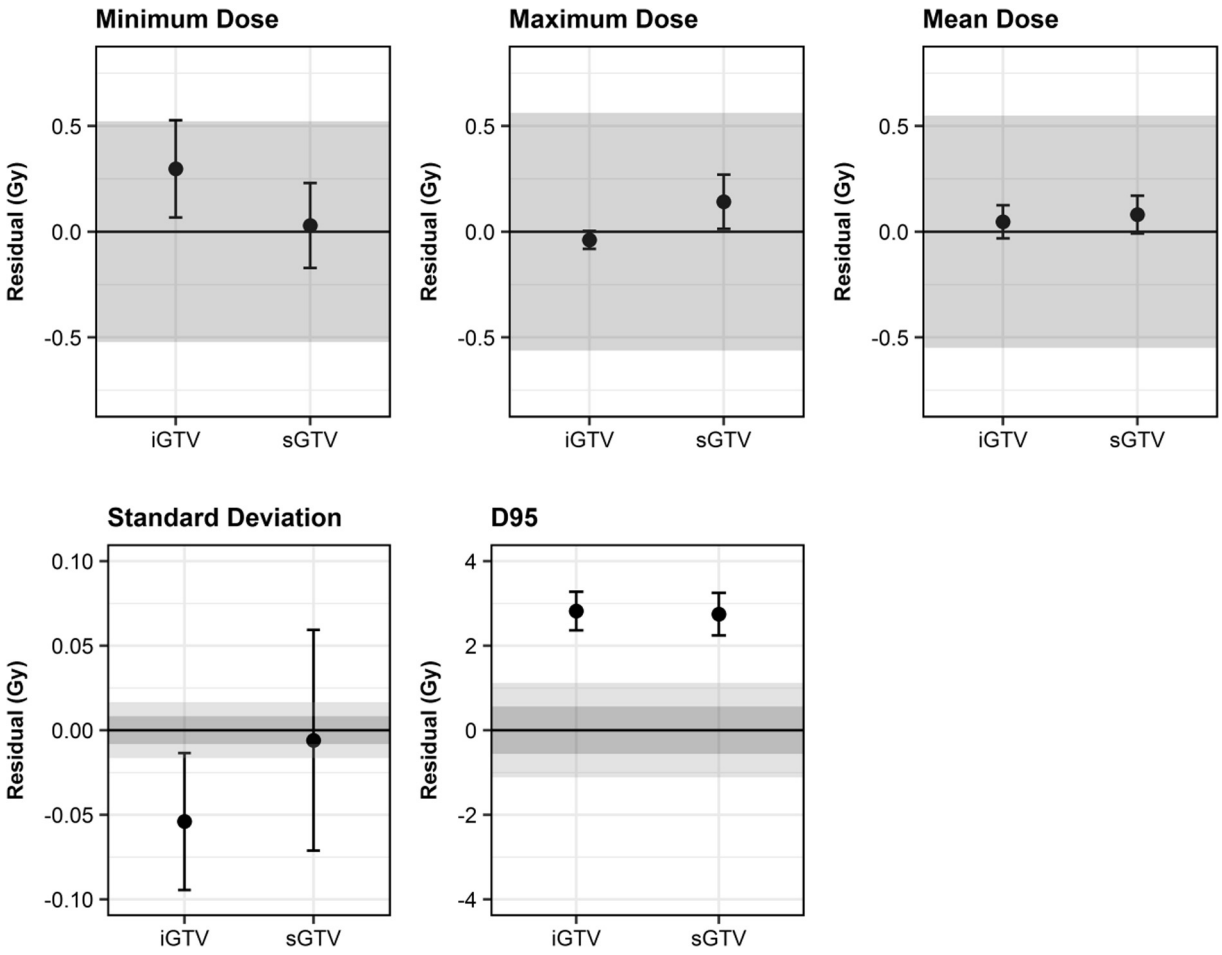

Fig. 2. Plots of the mean residual value for each dose metric as compared to the true GTV, calculated from all patients in the validation cohort. Error bars represent the 95% confidence interval of the mean residual, while the band represents the area of 'clinical indifference', calculated here as 1% of the lowest dose value obtained from the true GTV volumes, and additionally 2% of the true GTV value for the standard deviation and the D95.

Table 3
Random Forest Model Performance.

Model	AUC	Specificity	Sensitivity	Kappa
No GTV	0.606 ± 0.021	0.52	0.66	0.17
True GTV only	0.670 ± 0.011	0.56	0.74	0.29
Mean imputed GTVs	0.667 ± 0.006	0.42	0.82	0.25
RF imputed GTVs	0.691 ± 0.008	0.58	0.71	0.28
Combined true GTV and iGTV	0.689 ± 0.012	0.43	0.84	0.29
Combined true GTV and sGTV	0.696 ± 0.007	0.36	0.92	0.32

Resultant model performance values, where the mean and standard deviation of the AUC is based upon 20 repeats.

cohort size of lung cancer patients available for the development of radiotherapy outcome prediction models requiring parameters derived from GTV contours. As far as the authors are aware, this is the first study of this type.

The volumes predicted from the simple linear fits are all within the ±2 pixel volume tolerance, except in one lower lobe case. Both the upper and lower lobe tumour model fits have high R^2 values, but also relatively high RMSE values. The lower lobe model fit is of lower quality than the upper lobe model, which could be explained by the smaller number of patients available for modelling and/or by a greater variability in the motion magnitude for lower lobe patients. In any case all volume uncertainties are well within the magnitudes of inter-observer delineation variability reported in the literature. Steenbakkers et al. [24] found the mean observer variation over all patients included in their study to be 1.02 cm; Giraud et al. [25] observed mean differences of 2.1–3.1 cm; whilst Van de Steene et al. [26] found maximal differences in the GTV diameter of 4.2–8.4 cm. The maximum volume ratio over all cases is 1.69 for the linear fit. Typical V_{max}/V_{min} values from inter-observer delineation studies quoted in the literature, have been observed to be in the range of 1.1–10.1 [26–29].

Interestingly, contrary to existing knowledge of lung tumour motion, [18], the derived kernels were smaller in the cranial–caudal direction than in the left–right and anterior–posterior directions. This could be a result of the clinician not being blinded to the existing contours, and thus having limits to search for a GTV within; and differences in observer opinion regarding the inclusion of bifuricating regions, in combination with a possible underestimation of the motion due to the 3 mm slice interval. It should also be noted that the median tumour motion is only 4 mm peak-to-

peak in the cranial–caudal direction [30], corresponding to 2 mm on each side, which is similar to the slice interval. Regardless, the erosion method is superior to the linear fit, with all upper and lower cases meeting the 2 mm tolerance (12 cases meeting the 1 mm tolerance), and a maximum volume ratio of 1.32 (Table 2). The volume ratios for the erosion sGTV are also better than those for the iGTV – with mean values closer to 1, and lower standard deviations for both upper and lower lobe cases.

The closest point method, used to assess the shape of the contours, found relatively large maximum closest point distances (>2 cm) in 2 out of the 15 validation patients, both of which appear in the lower lobe cohort. These shape deviations correspond to differences in observer opinion as to whether or not to include some bifurcating regions within the iGTV/true GTV delineation at the site of gross disease. Additionally, the closest point distance is the discrete case of the Hausdorff distance metric and can over-estimate comparable distances as measurements may not be normal to contours. As discussed above, these errors remain within the bounds of observer errors reported in the literature. No statistical difference was observed between the closest point measurements of the sGTV and the iGTV. Similarly, while the Dice Similarity Coefficients (Table 2) show good agreement between the true and synthetic GTVs contours, no significant difference between the DSC for the sGTV and for the iGTV was found for either the upper or lower lobe cases.

The model developed using a cohort consisting only of patients with an sGTV value was found to successfully predict the outcome of a test cohort consisting only of patients with a true GTV value, and vice versa. This therefore further strengthens the argument that there is good agreement between the true and synthetic volumes. Furthermore, the model built with a combined sGTV and true GTV cohort performed significantly better against an external validation dataset than all other models, including the model built from true GTVs alone, models of equal size but with imputed GTV data, or the model built using GTVs and iGTVs. This shows the potential of using the sGTV, though the difference in performance is small. The main benefit of using sGTVs over iGTVs is that it allows truly large datasets to be analysed without having to re-delineate the GTVs and avoids a potential bias that could occur when GTVs and iGTVs are mixed.

The external validation data set used for modelling contained GTVs that were defined on 3D CT scans. The use of 3D CT can introduce distortions to the tumour volume definition, either over- or under-estimation, due to the tumour being captured in different phases of the breathing cycle in different slices [31], and therefore could affect the results. With the introduction of multi-slice helical scanners, however, the tumour is much more likely to be captured in a single or a few breathing phases, minimising distortions in all but a few cases with the largest tumour volumes and/or the quickest breathing cycles. It is therefore most likely that the external validation cohort will contain a few outliers, but that the majority of cases would have minimal distortion and therefore the images used for target delineation would have been representative of the tumour volume.

Our results show that the developed erosion sGTV provides a superior volume estimate to both the linear fit sGTV and the iGTV, and is equivalent to the iGTV in terms of its shape. This is demonstrated by the volume ratios – which are increasingly improved as compared to the iGTV by the use of the linear fit and the erosion method – as well as the modelling results, which found the model containing the sGTV to be superior to all others developed. At present, only the volume is of interest for prediction modelling, for which the presented synthesis method is effective. However, if any kind of shape analysis is required, for example, for radiomics studies, then another, more sophisticated method will need to be developed.

Colloquially, the iGTV is often referred to as the ITV, despite its definition differing from that of the ICRU. For data sharing regimes, it will be vital that the terminology is clarified. Depending on how each centre defines and outlines these volumes, it might be that different kernels for the erosion are required; however, the methodology developed here should still be applicable. We recommend in any case all results to be confirmed in each centre's own data.

Conclusion

We demonstrate a purposely simple methodology, suitable for automated application to large databases of radiotherapy patients, to estimate GTV volume from existing iGTV contours. The uncertainty in the volume estimates is within reported inter-observer variabilities; and a cohort utilising the synthesised volumes was superior to others built using the true GTV only or utilising the iGTV. The technique provides a robust estimation of volume, but the effect of spiculation on metrics that might be more sensitive to shape (e.g. some radiomics measures) remains to be investigated. The erosion kernels used were developed against iGTVs defined on patients' MIPs and may need changing where different contouring protocols are used.

Conflict of interest statement

Nothing to disclose.

Acknowledgements

This study was funded by the Manchester Cancer Research UK Major Centre Award.

References

[1] El Naqa I, Li R, Murphy M. Machine learning in radiation oncology, theory and applications. Springer; 2015.
[2] Dehing-Oberije C, Yu S, De Ruysscher D, Meersschout S, Van Beek K, Lievens Y, et al. Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys 2009;74:355–62. https://doi.org/10.1016/j.ijrobp.2008.08.052.
[3] Dehing-Oberije C, De Ruysscher D, Petit S, Van Meerbeeck J, Vandecasteele K, De Neve W, et al. Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. Radiother Oncol 2010;97:455–61. https://doi.org/10.1016/j.radonc.2010.09.028.
[4] 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;42:2421–30. https://doi.org/10.1118/1.4915284.
[5] Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, et al. "Rapid Learning health care in oncology" – an approach towards decision support systems enabling customised radiotherapy. Radiother Oncol 2013;109:159–64. https://doi.org/10.1016/j.radonc.2013.07.007.
[6] Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology–multifactorial decision support systems. Nat Rev Clin Oncol 2013;10:27–40. https://doi.org/10.1038/nrclinonc.2012.196.
[7] Price G, van Herk M, Faivre-Finn C. Data mining in oncology: the ukCAT project and the practicalities of working with routine patient data. Clin Oncol 2017:10–3. https://doi.org/10.1016/j.clon.2017.07.011.
[8] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward Michael G, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338. https://doi.org/10.1136/bmj.b2393.
[9] International Commission on Radiation Units and Measurements. ICRU Report 50: Prescribing, Recording and Reporting Photon Beam Therapy; 1993.
[10] Oberije C, De Ruysscher D, Houben R, van de Heuvel M, Uyterlinde W, Deasy JO, et al. A validated prediction model for overall survival from stage III non-small cell lung cancer: toward survival prediction for individual patients. Int J Radiat Oncol Biol Phys 2015:1–10. https://doi.org/10.1016/j.ijrobp.2015.02.048.
[11] Reymen B, Van Loon J, van Baardwijk A, Wanders R, Borger J, Dingemans A-MC, et al. Total gross tumor volume is an independent prognostic factor in patients treated with selective nodal irradiation for stage I to III small cell lung cancer.

Int J Radiat Oncol Biol Phys 2013;85:1319–24. https://doi.org/10.1016/j.ijrobp.2012.10.003.
[12] Tsujino K, Hirota S, Endo M, Obayashi K, Kotani Y, Satouchi M, et al. Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer. Int J Radiat Oncol Biol Phys 2003;55:110–5. https://doi.org/10.1016/S0360-3016(02)03807-5.
[13] van Loon J, Siedschlag C, Stroom J, Blauwgeers H, van Suylen R-J, Knegjens J, et al. Microscopic disease extension in three dimensions for non – small-cell lung cancer: development of a prediction model using pathology-validated positron emission tomography and computed tomography features. Int J Radiat Oncol 2012;82:448–56. https://doi.org/10.1016/j.ijrobp.2010.09.001.
[14] 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. https://doi.org/10.1038/ncomms5006.
[15] International Commission on Radiation Units and Measurements; ICRU Report 62: Prescribing, Recording and Reporting Photon Beam Therapy (supplement to ICRU Report 50); 1999.
[16] Zamora DA, Riegel AC, Sun X, Balter P, Starkschall G, Mawlawi O, et al. Thoracic target volume delineation using various maximum-intensity projection computed tomography image sets for radiotherapy treatment planning. Med Phys 2010;37:5811–20. https://doi.org/10.1118/1.3504605.
[17] Seppenwoolde Y, Shirato H. Precise and real-time measurement of 3D tumor motion in lung due to breathing and heartbeat, measured during radiotherapy. Int J Radiat Oncol Biol Phys 2002;53:822–34.
[18] Sonke J-J, Lebesque J, van Herk M. Variability of four-dimensional computed tomography patient models. Int J Radiat Oncol Biol Phys 2008;70:590–8. https://doi.org/10.1016/j.ijrobp.2007.08.067.
[19] International Commission on Radiation Units and Measurements. ICRU Report 42: Use of Computers in External Beam Radiotherapy Procedures with High-Energy Photons and Electrons; 1987.
[20] Kuhn M, Weston S, Keefer C, Engelhardt A, Cooper T, Mayer Z, et al. caret: Classification and Regression Training; 2016.
[21] Ishwaran H, Kogalur UB. Random Forests for Survival, Regression and Classification (RF-SRC); 2016.
[22] Ishwaran H, Kogalur U. Random survival forests for R. RNews 2007;7:25–31.
[23] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841–60. https://doi.org/10.1214/08-AOAS169.
[24] Steenbakkers RJHM, Duppen JC, Fitton I, Deurloo KEI, Zijp L, Uitterhoeve ALJ, et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a "Big Brother" evaluation. Radiother Oncol 2005;77:182–90. https://doi.org/10.1016/j.radonc.2005.09.017.
[25] Giraud P, Elles S, Helfre S, De Rycke Y, Servois V, Carette M, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. Radiother Oncol 2002;62:27–36.
[26] Van De Steene J, Linthout N. Definition of gross tumor volume in lung cancer: inter-observer variability. Radiother Oncol 2002;62:37–49.
[27] Vorwerk H, Beckmann G, Bremer M, Degen M, Dietl B, Fietkau R, et al. The delineation of target volumes for radiotherapy of lung cancer patients. Radiother Oncol 2009;91:455–60. https://doi.org/10.1016/j.radonc.2009.03.014.
[28] Weiss E, Hess CF. The impact of gross tumor volume (GTV) and clinical target volume (CTV) definition on the total accuracy in radiotherapy theoretical aspects and practical experiences. Strahlentherapie Und Onkol Organ Der Dtsch Röntgengesellschaft 2003;179:21–30. https://doi.org/10.1007/s00066-003-0976-5.
[29] Bowden P, Fisher R, Mac Manus M, Wirth A, Duchesne G, Millward M, et al. Measurement of lung tumor volumes using three-dimensional computer planning software. Int J Radiat Oncol Biol Phys 2002;53:566–73.
[30] Peulen H, Belderbos J, Rossi M, Sonke JJ. Mid-ventilation based PTV margins in Stereotactic Body Radiotherapy (SBRT): a clinical evaluation. Radiother Oncol 2014;110:511–6. https://doi.org/10.1016/j.radonc.2014.01.010.
[31] Balter J, Ten Haken R, Lawrence T. Uncertainties in CT-based radiation therapy treatment planning associated with patient breathing. Int J 1996;36:167–74.