Radiotherapy and Oncology 130 (2019) 2–9

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original article

Vulnerabilities of radiomic signature development: The need for safeguards

Mattea L. Welch [a,f,i], Chris McIntosh [e,f,i], Benjamin Haibe-Kains [a,c,i,j], Michael F. Milosevic [b,e,i], Leonard Wee [g], Andre Dekker [g], Shao Hui Huang [b,i], Thomas G. Purdie [b,e,f,i], Brian O'Sullivan [b,i], Hugo J.W.L. Aerts [h], David A. Jaffray [a,b,d,e,f,i,*]

[a] Department of Medical Biophysics, University of Toronto; [b] Department of Radiation Oncology, University of Toronto; [c] Ontario Institute of Cancer Research, Toronto; [d] IBBME, University of Toronto; [e] Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto; [f] The Techna Institute for the Advancement of Technology for Health, Toronto, Canada; [g] Department of Radiation Oncology (MAASTRO), GROW Research Institute, Maastricht University, the Netherlands; [h] Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, USA; [i] Princess Margaret Cancer Centre, University Health Network; and [j] Vector Institute, Toronto, Canada

ARTICLE INFO

Article history:
Received 29 May 2018
Received in revised form 17 October 2018
Accepted 23 October 2018
Available online 8 November 2018

Keywords:
Radiomics
Signature development
Safeguards
Volume
Lung cancer
Head and neck cancer

ABSTRACT

Purpose: Refinement of radiomic results and methodologies is required to ensure progression of the field. In this work, we establish a set of safeguards designed to improve and support current radiomic methodologies through detailed analysis of a radiomic signature.

Methods: A radiomic model (MW2018) was fitted and externally validated using features extracted from previously reported lung and head and neck (H&N) cancer datasets using gross-tumour-volume contours, as well as from images with randomly permuted voxel index values; i.e. images without meaningful texture. To determine MW2018's added benefit, the prognostic accuracy of tumour volume alone was calculated as a baseline.

Results: MW2018 had an external validation concordance index (c-index) of 0.64. However, a similar performance was achieved using features extracted from images with randomized signal intensities (c-index = 0.64 and 0.60 for H&N and lung, respectively). Tumour volume had a c-index = 0.64 and correlated strongly with three of the four model features. It was determined that the signature was a surrogate for tumour volume and that intensity and texture values were not pertinent for prognostication.

Conclusion: Our experiments reveal vulnerabilities in radiomic signature development processes and suggest safeguards that can be used to refine methodologies, and ensure productive radiomic development using objective and independent features.

© 2018 The Author(s). Published by Elsevier B.V. Radiotherapy and Oncology 130 (2019) 2–9 This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Cancer imaging features used to inform medical decisions are generated by experienced radiologists, often involving qualitative and experiential interpretation [1,2]. However, utilization of quantified patient imaging data for pattern recognition has recently increased. Radiomics involves the automated extraction of imaging features for use in multivariate predictions models and has demonstrated promise in defining predictive and prognostic factors [3] for disease relapse and mortality after treatment [4–7], and biological correlates [8,9]. The largest study to date was published in 2014 by Aerts et al. [7], and showed the prognostic power of a radiomic signature for overall survival in lung, and head and neck (H&N) cancer. Radiomic features corresponding to shape, texture and image intensity were extracted from within the gross-tumour-volume (GTV) of 1019 pre-treatment CT images. A four-feature signature (First order-Energy, Shape-Compactness, Texture-Gray level non-uniformity (GLNU), and Wavelet HLH-GLNU) was defined by focusing on the most stable features for prognostication in a lung dataset, and validated using independent lung and H&N patient cohorts.

Aerts et al.'s study [7] provided detailed descriptions of their methods and data, permitting usage of their signature by multiple groups. Leijenaar et al. [10] externally validated the model and signature using a H&N dataset from The Princess Margaret Cancer Centre. Grossman et al. [8] and Vallières et al. [6] also successfully used the model and signature for lung and H&N prognostication and outcome predictions. However, the original model was developed using in-house Matlab code [11], which was not openly released and thus limited reproducibility studies to groups with access to this software.

Accessibility of software is a substantial concern since radiomic platforms are not interchangeable [12]; therefore, reducing generalization and potential impact of published studies and models.

* Corresponding author at: 190 Elizabeth St. RFE, 1S-411, University Health Network, Toronto, Ontario M5G 2M9, Canada.

E-mail address: david.jaffray@rmp.uhn.ca (D.A. Jaffray).

https://doi.org/10.1016/j.radonc.2018.10.027
0167-8140/© 2018 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Thus to provide the community with an open and modular platform, van Griethuysen et al. recently released PyRadiomics [13], an open-source implementation of the methods contained in the original radiomic Matlab code [7]. Open-source software that complies with accepted standards will improve generalization of published studies and models, while promoting further independent development and evaluation of published models. The availability of open-source software, like PyRadiomics, and the original lung training dataset [14] used by Aerts [7] permits further investigation of the prognostic and predictive value of these radiomic features.

Through a detailed analysis of a radomic model this manuscript describes broad vulnerabilities and critical learnings in the application of established radiomic methodologies for the development and testing of radiomic signatures. These were observed using the original lung dataset [14], and validated using the H&N dataset used by Leijenaar et al. [10]. Safeguards to defend against underlying dependences revealed through exhaustive testing of model and feature multicollinearity are reported here. These are designed for adoption by the community to improve the interpretability and accuracy of future radiomic signatures and models.

Materials and methods

Our analysis of a refitted model and its corresponding features deconstructs the prognostic accuracy by working in a "backwards" manner. By beginning our analysis at the final model and moving towards the univariate features we can suggest safeguards that will protect future radiomic developments from underlying dependencies and feature multicollinearity. This section details the datasets and methods used for our analysis.

Datasets

This study used two datasets: (1) The lung cancer dataset [14] (Lung1) used by Aerts et al. [7] was obtained by the MAASTRO Clinic. Pre-treatment CT images and GTV contours were available for all original 422 patients; however, one patient had undergone surgical resection of their lung and was not used for this study. Disease status details for the 421 patients can be found in (Supplementary Table 1). The 421 patients were treated with induction concurrent chemoradiotherapy, radiotherapy alone, or chemotherapy followed by radiotherapy. Lung1 was used for model training; (2) The H&N cancer dataset (H&N1) reported by Leijenaar [10] (previously named PMH1) was obtained from The Princess Margaret Cancer Centre with institutional research board approval to validate the fitted model. H&N1 comprised radiotherapy planning CT images and primary GTVs for 527 oropharyngeal squamous cell carcinoma patients treated with concurrent chemoradiotherapy or radiotherapy alone between 2005 and 2010 at The Princess Margaret Cancer Centre (Supplementary Table 2). GTVs were contoured by attending radiation oncologists based on available clinical-radiological evidence of disease extent, and used for IMRT treatment. Diagnostic magnetic resonance imaging (MRI) was often fused with planning CT to aid target delineation. Radiotherapy target volumes, including GTV were peer-reviewed (slice-by-slice comparison of planning CT and MRI) during weekly H&N Radiation Oncology Quality Assurance Rounds as part of routine clinical practice. Patients with multiple GTVs were excluded from both datasets, this resulted in a reduction of patients in the H&N1 dataset from 542 to 527.

All methods and experimental protocols were in accordance with relevant guidelines and regulations. The study was approved by the Institutional Research Ethic board (IRB). Informed consent of the subjects was waived by our IRB since this is a retrospective review; a waiver would not affect the rights and welfare of the study subjects.

Radiomic feature extraction

PyRadiomics [13] was used to extract features from Lung1 and H&N1 GTVs. We did not select new features, and instead used the four features with the same name as those described previously by Aerts et al. [7] (Supplementary Table 3): (1) Energy, from the first order feature class of PyRadiomics, calculates the overall density of the tumour volume. This feature is also commonly referred to as an intensity histogram-based uniformity feature [15]; (2) Shape compactness, compactness of tumour volume relative to sphere. We chose to use Compactness 1 in this work which has been recently used in other reproducibility studies [6,10]; (3) Gray level non-uniformity (GLNU, defined as GLN in PyRadiomics documentation), intra-tumour heterogeneity from gray level run length matrices. A gray level run length matrix is calculated for all directions in the 3D imaging volume, and a corresponding GLNU feature is calculated for each direction. The resulting GLNU values for the 13 different directions are then averaged together to obtain the final GLNU features; and (4) HLH-GLNU, the same as (3), but after wavelet decomposition of the original image along the x, y and z-axes using a coif wavelet (Fig. 1). For consistency with Aerts et al. [7] we did not interpolate patient images to uniform voxel sizes in all three dimensions or between patients, and used a fixed bin width of 25 for discretization of images and wavelets with a starting bin edge of 0, the default for PyRadiomics. More detailed explanations of feature calculations can be found in Pyradiomics documentation [13] and the Image Biomarker Standardization Initiative [15].

Additional tumour volume (in mm^3, the default of PyRadiomics) and tumour volume normalized features were also extracted. Volume normalized feature calculation methods were added to PyRadiomics, and defined by adding a term to the original feature equation that normalized by the total number of voxels in the GTV [16] (Supplementary Table 3). These features were used later to test for feature volume dependencies.

MW2018 model fitting and validation with and without voxel randomization

Due to possible implementation differences [12] between the original feature definitions in Aerts et al. [7], and the open-source toolbox, PyRadiomics, we refit the Cox model-weights using Lung1 imaging features and the Survival R package (version 2.41–3). Our model was fitted to predict overall survival and is henceforth referred to as the MW2018 model to represent the first author's initials and year of publication.

The MW2018 model was externally validated using H&N1 features. Concordance index (c-index) values were calculated using prognostic index (PI) values (Fig. 1), where a high PI corresponded to low expected survival. PI is a linear predictor of the Cox model that expresses the risk for an event, and is output by the model as the sum of the products of the model coefficients multiplied by their corresponding features. We used the survcomp R package (version 1.28.4) [17] to compute the point estimate of c-index, its 95% confidence interval (CI) and p-value (two-sided, $\alpha = 0.05$) using the noether approach [18]. Additionally, Kaplan–Meier curves were generated for MW2018 PI values, volume and, univariate features to describe overall survival since date of diagnosis. Log-rank tests compared survival distributions of high and low risk cohorts.

To examine MW2018's reliance on intensity features we tested it on simulated images devoid of meaningful intensity signals. Image intensity values of each patient's CT were shuffled by randomly permuting voxel index values across the entire 3D dataset (Fig. 2). The newly voxel-randomized image, along with the

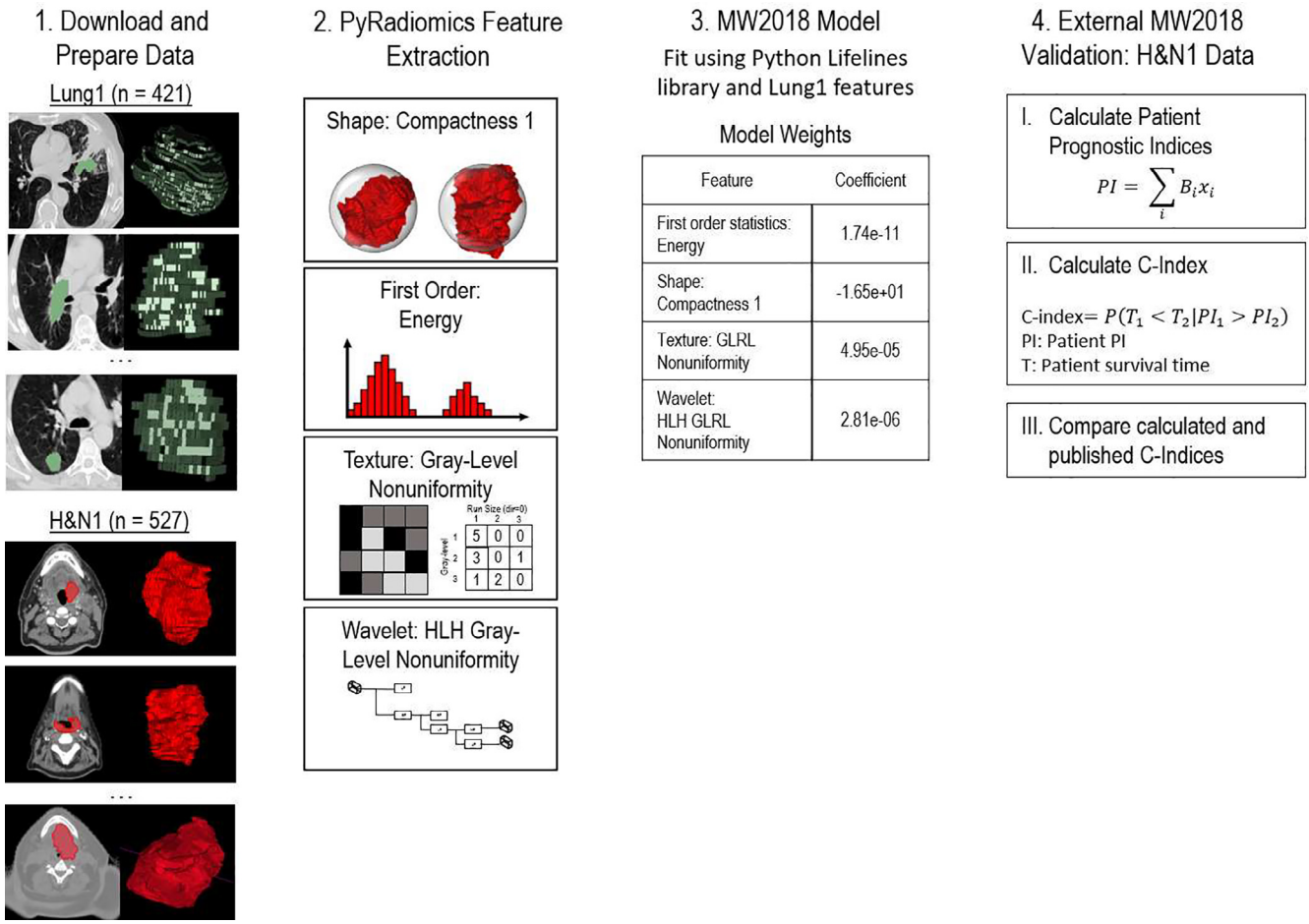

Fig. 1. Workflow for training and usage of MW2018 Model: (1) CT images and GTVs are selected for Lung1 and H&N1 datasets. The selected DICOM data are converted to Python compatible file types; (2) Signature features are extracted using Pyradiomics; (3) Lung1 signature features are used to train a Cox model; (4) MW2018 model coefficients are used with extracted H&N1 features to calculate patient specific prognostic indices (PI). The resulting concordance index (c-index) is calculated.

unaltered clinical GTV contour, were then used for radiomic feature extraction. Ten sets of signature features were extracted from 10 separate Lung1 and H&N1 voxel-randomized datasets (Fig. 2). The c-index was calculated for all voxel-randomized univariate signature features for all ten sets of features. Additionally, MW2018 was applied to H&N1 extracted voxel-randomized features, and the c-index was calculated for the resulting PI values.

Analysis of signature tumour volume dependence and multicollinearity

The c-index of MW2018 was statistically (two-sided) compared to tumour volume using the compareC R package (version 1.3.1) [19] in the H&N1 dataset using unaltered and voxel-randomized images; statistically significant differences in c-index values were indicated by the calculated p-value. Correlation of tumour volume to signature features and MW2018 PIs was tested using Spearman rank correlation values (ρ), which was selected to ensure analysis of all potential monotonic correlations. Highly correlated features were normalized for volume and ρ was recalculated. Univariate c-index and log-rank test p-values for features before and after normalization were compared to determine dependence on volume.

Independent prognostic information and interactive effects of features weakly correlated with volume were determined by fitting two Cox models and analysing covariant coefficient p-values. Model 1 was built using volume and a weakly correlated feature. Model 2 was built using volume, a weakly correlated feature, and the interactive term between the two as covariates. The feature vectors were normalized to unit vectors prior to fitting the Cox models by dividing each element (i.e. feature) by the vector magnitude. This was achieved using Eq. (1) where k is a feature vector and $|k|$ is the norm, or magnitude, of the vector. The vector norm is calculated by taking the square root of the sum of the squared vector values. By normalizing the feature vectors prior to Cox model fitting, the calculated hazard ratios better reflected relative importance of the features.

$$\hat{k} = \frac{k}{|k|} \tag{1}$$

Extent of coefficient standard error inflation due to signature feature multicollinearity was explored in the training dataset using Variance Inflation Factors (VIF) [20,21]. VIFs provide a measurement of how estimated regression coefficient variance is increased by collinearity, and were calculated for each of the previously selected signature features in Lung1 using Eq. (2). For a given feature, k, a linear regression model was fit to predict k using the other three features. R_k^2 is the coefficient of determination for the linear regression model fit without feature k.

$$VIF_k = \frac{1}{1 - R_k^2} \tag{2}$$

The square root of a VIF indicates how much larger the standard error is in comparison to the case in which the feature was uncorrelated with other features in the model.

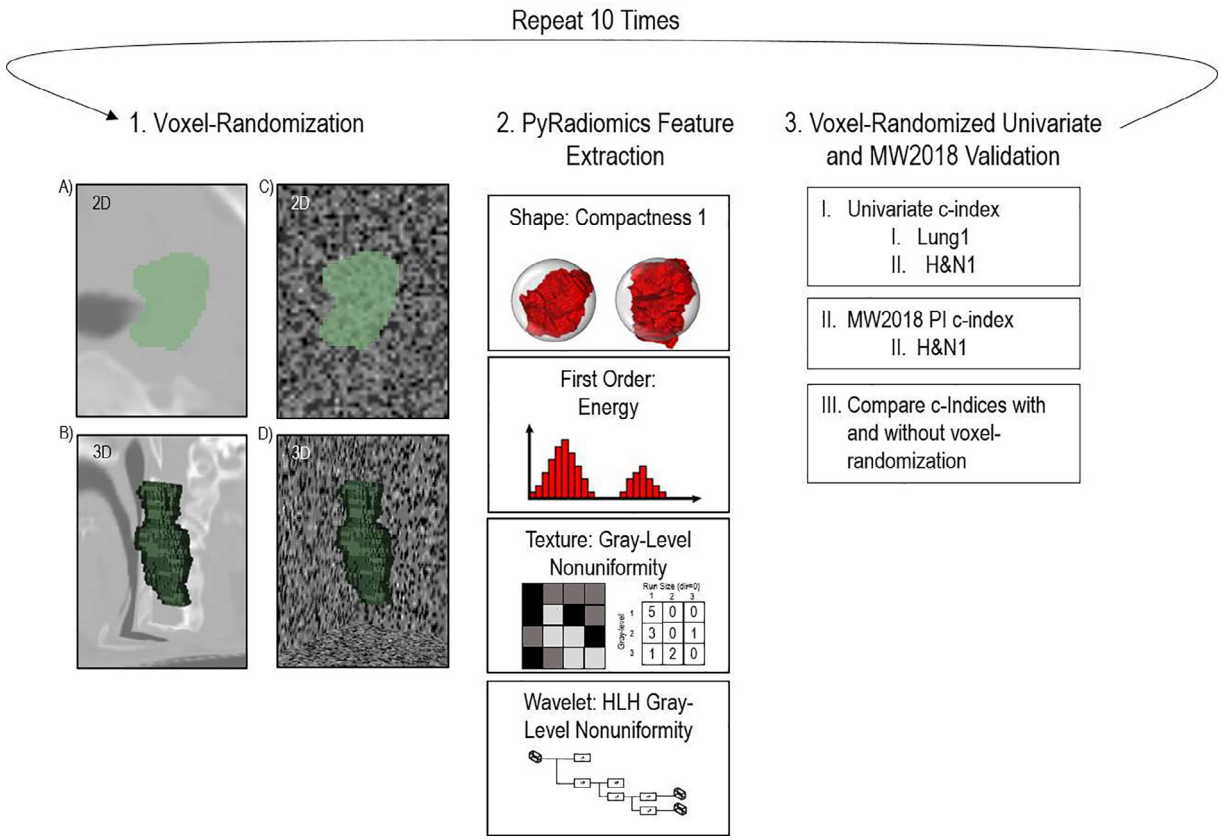

Fig. 2. Workflow for analysing MW2018's reliance on intensity features: (1) CT image voxel locations are randomized in all Lung1 and H&N1 datasets. (A) 2D axial image slice with contour before voxel-randomization; (B) 3D image volume with a rendered contour before voxel-randomization; (C) 2D voxel-randomized axial image slice with contour; (D) 3D voxel-randomized image volume with rendered contour; (2) The unaltered contours and their corresponding voxel-randomized images are used for feature extraction with PyRadiomics; (3) Univariate c-index values are calculated for signature features in both datasets. Our MW2018 model is applied to the signature features extracted from H&N1. The c-index values calculated with and without voxel-randomized images are compared to determine if the features and MW2018 are as prognostic when using images without meaningful intensity values.

Results

Model validation with and without Voxel-Randomized images

The MW2018 model had a c-index of 0.59 (CI = 0.55–0.63 *p*-value < 0.05) (Table 1) in the Lung1 data when predicting overall survival. MW2018 yielded a c-index of 0.64 (CI = 0.60–0.68, *p*-value < 0.05) in the H&N1 dataset (Table 1 and Fig. 3d). Univariate feature data for energy, compactness 1, GLNU, and HLH-GLNU had c-index of 0.57, 0.45, 0.61 and 0.60, respectively in Lung1, and 0.61, 0.42, 0.64, and 0.63 (CI and *p*-values in Table 1), respectively in H&N1. Kaplan–Meier curves for the four univariate features are in Supplementary Figs. 3 and 4.

The same MW2018 model had c-index values of 0.60 and 0.64 when applied to voxel-randomized Lung1 and H&N1 features, respectively (Fig. 3b and e); the 10 different voxel- randomizations resulted in c-index standard deviations less than 0.006 (Supplementary Fig. 2). Individual features of energy, compactness 1, GLNU, and HLH-GLNU had c-index = 0.60, 0.45, 0.60, 0.60, respectively in Lung1 and c-index = 0.63, 0.42, 0.64, 0.64, respectively

Table 1
c-Index and *p*-values from original and voxel-randomized Lung1 and H&N1 datasets. 95% CI are shown for point-estimated c-index values. Standard deviations are reported for c-index values calculated across multiple runs, namely the application of MW2018 to the 10 randomized datasets, and the univariate calculations on the 10 randomized datasets.

			Lung1		H&N1	
	Feature	Image	c-Index (95% CI)	*p*-Value	c-Index (95% CI)	*p*-Value
Univariate	Energy	Original	0.57 (0.53–0.61)	<0.05	0.61 (0.57–0.65)	<0.05
	Compactness 1		0.45 (0.41–0.49)	<0.05	0.42 (0.38–0.47)	<0.05
	GLNU		0.61 (0.57–0.64)	<0.05	0.64 (0.59–0.68)	<0.05
	HLH GLNU		0.60 (0.56–0.64)	<0.05	0.63 (0.59–0.68)	<0.05
	Volume		0.60 (0.56–0.64)	<0.05	0.64 (0.59–0.68)	<0.05
	Energy	Randomized	0.60 ± 0.006	<0.001	0.63 ± 0.0004	<0.05
	Compactness 1		0.45 ± 0.0	0.027	0.42 ± 0.0	<0.05
	GLNU		0.60 ± 0.006	<0.001	0.64 ± 0.0003	<0.05
	HLH GLNU		0.60 ± 0.006	<0.001	0.64 ± 0.0003	<0.05
	Normalized Energy	Original	0.43 (0.39–0.47)	<0.05	0.46 (0.42–0.50)	0.058
	Normalized GLNU		0.58 (0.54–0.62)	<0.05	0.56 (0.52–0.61)	<0.05
	Normalized HLH GLNU		0.57 (0.53–0.61)	<0.05	0.57 (0.52–0.61)	<0.05
MW2018 Application		Original	0.59 (0.55–0.63)	<0.05	0.64 (0.60–0.68)	<0.05
		Randomized	0.60 ± 0.006	N/A	0.64 ± 0.0003	N/A

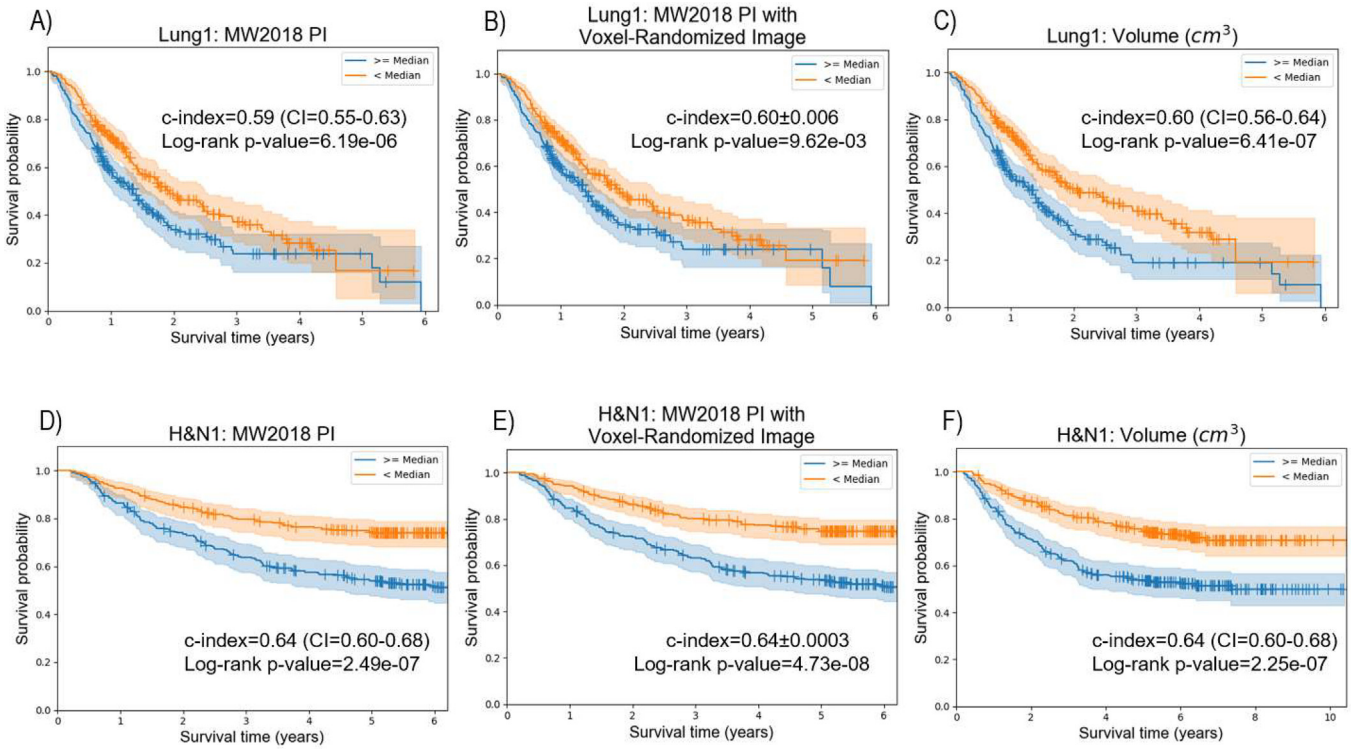

Fig. 3. Kaplan–Meier curves displaying radiomic signature performance during stratification of patients into high and low risk groups: (A) good performance in Lung1 with training concordance index (c-index) of 0.59 (CI = 0.55–0.63). The log-rank test of the Kaplan–Meier curve had a *p*-value = 6.19e−06; (B) good performance in randomized Lung1 image volumes with a c-index across 10 runs of 0.60 ± 0.006 and an example log-rank test *p*-value = 9.62e−03; (C) good performance of tumour volume in Lung1 with c-index = 0.60 (CI = 0.56–0.64) and a log-rank test *p*-value = 6.41e−07; (D) good performance MW2018 in H&N1 with c-index = 0.64 (CI = 0.60–0.68) and a log-rank test *p*-value = 2.83e−10; E) good performance in 10 randomized iterations with H&N1 image volumes with c-index = 0.64 ± 0.0003 and an example log-rank test *p*-value = 4.73e−08; (F) good performance of tumour volume in H&N1 with c-index = 0.64 (CI = 0.59–0.68) and a log-rank test *p*-value = 2.25e−7.

in H&N1; standard deviation was less than 0.006 for all univariate feature CIs (Supplementary Fig. 2). Kaplan–Meier curves for the four univariate features extracted from voxel-randomized images are in Supplementary Figs. 5 and 6.

Signature tumour volume dependence analysis and multicollinearity

Tumour volume alone had a c-index of 0.60 (CI = 0.56–0.64, *p*-value < 0.05) and 0.64 (CI = 0.59–0.68, *p*-value < 0.05) (Fig. 3c and f) in Lung1 and H&N1, respectively. When the c-index of univariate volume is compared to the c-index of MW2018 with unaltered H&N1 images we cannot conclude that MW2018 is more prognostic than univariate volume (*p*-value = 0.90). Additionally, based on our data and experiments, we cannot conclude that the model applied to voxel-randomized images is significantly different from univariate volume (average *p*-value = 0.95).

We calculated spearman rank values for tumour volume versus multivariate MW2018 PI, energy, GLNU, and HLH-GLNU (Table 2). The spearman rank values of energy, GLNU and HLH-GLNU all changed after volume normalization (Table 2). c-index values of volume-normalized features were also reduced in both Lung1 and H&N1 (c-index values and *p*-values in Table 1). Kaplan–Meier curves for univariate features before and after volume normalization are in Supplementary Figs. 3 and 4. The VIF values for energy, compactness 1, GLNU, and HLH-GLNU in Lung1 are 1.94, 1.41, 56.56, and 62.68, respectively (Supplementary Table 4), indicating high multicollinearity between GLNU and HLH-GLNU.

Compactness 1 and tumour volume had spearman rank values of $\rho = -0.50$ and -0.38, in Lung1 and H&N1, respectively. Tumour volume, compactness and an interaction term between the two were tested using multivariate Cox models, as summarized in Table 3.

Table 2
Spearman rank correlation values comparing tumour volume to patient MW2018 PIs and individual radiomic signature features before and after tumour volume normalization in Lung1 and H&N1.

	Feature name	Spearman rank value	
		Pre volume normalization	Post volume normalization
H&N1	First Order: Energy	0.83	0.28
	Shape: Compactness 1	−0.38	N/A
	Texture: GLNU	0.95	−0.51
	Wavelet: HLH GLNU	0.99	−0.28
	MW2018 PI	0.79	N/A
Lung1	First Order: Energy	0.76	−0.49
	Shape: Compactness 1	−0.50	N/A
	Texture: GLNU	0.97	0.58
	Wavelet: HLH GLNU	0.99	0.49
	MW2018 PI	0.87	N/A

Interaction terms were not significant in either Lung1 or H&N1 (*p*-value = 0.91 and 0.83 respectively), indicating that the prognostic effect of compactness does not depend on volume significantly, and vice versa. When tumour volume and compactness without the interaction term were tested in a Cox model, both provided independent prognostic information in H&N1 (Volume: HR = 2.21e + 05, *p*-value < 0.001; Compactness: HR = 1.08e−09, *p*-value = 0.02) but only volume was significant in Lung1 (Volume: HR = 9.30e+02, *p*-value < 0.001; Compactness: HR = 5.47e−04, *p*-value = 0.21).

Discussion

Refinement of methodology and increased transparency are integral to the future of radiomics. In this work, a Cox model was

Table 3
Coefficients for Cox model with weakly correlated covariates. Interactive effects are tested for by multiplying two covariates of interest (represented by *). Significant *p*-values ($\alpha = 0.05$) are highlighted. All feature vectors have been scaled to unit vectors.

Model		Feature	Coefficient (95% CI)	Hazard Ratio (95% CI):	*p*-Value	c-Index
H&N1	1	Volume	1.23e+01 (9.38e+00–1.52e+01)	2.21e+05 (1.19e+04–4.12e+06)	<0.001	0.64
		Compactness 1	-2.06e+01 (−3.52e+01 to −6.10e+00)	1.08e−09 (5.24e−16–2.23e−03)	0.02	
	2	Volume	1.03e+01 (−5.28e+00–2.58e+01)	2.88e+04 (5.06e−03–1.63e+11)	0.28	0.64
		Compactness 1	−2.26e+01 (−4.32e+01–−1.97e+00)	1.52e−10 (1.67e−19–1.39e−01)	0.07	
		Volume* Compactness 1	4.93e+01 (−3.19e−02 to 4.18e+02)	2.67e+21 (2.29e−139–3.11e+181)	0.83	
		Volume	6.84e+00 (4.29e+00–9.38e+00)	9.30e+02 (7.29e+01–1.19e+04)	<0.001	0.60
Lung1	1	Compactness 1	−7.51e+00 (−1.73e+01–2.27e+00)	5.47e−04 (3.10e−08–9.65e+00)	0.21	
		Volume	6.02e+00 (−6.30e+00–1.83e+01)	4.13e+02 (1.85e−03–9.26e+07)	0.42	
		Compactness 1	−8.01e+00 (−2.03e+01–4.25e+00)	3.33e−04 (1.58e−09–7.02e+01)	0.28	0.60
	2	Volume* Compactness 1	1.69e+01 (−2.33e−02–2.67e−02)	2.10e+07 (4.24e−102–1.04e+116)	0.91	

fitted that had a prognostic accuracy of 0.64 when externally validated, similar to previous publications utilizing the same signature [6–8],10,22]. However, it was determined that tumour volume alone was as prognostic as the model in H&N1 (c-index = 0.64 (CI = 0.59–0.68) vs. c-index = 0.64 (CI = 0.60–0.68), *p*-value = 0.90), and was embedded in three of the four signature features, suggesting a lack of feature independence (Tables 1 and 2). In the course of the work described here, several "safeguard" tests were formulated to guide the development of future radiomic signatures and models to ensure productive model development.

Radiomic models and signatures embedded or confounded by underlying features are a concern when drawing conclusions between features and disease relationships. Underlying dependencies may also impede prognostic and predictive accuracy since resulting models will have been fit using multiple surrogates of the same feature. With the above concerns in mind, we recommend the following safeguards be followed in future radiomic studies:

(1) Open-source software, such as PyRadiomics, should be used to increase development accountability and facilitate inter-institutional research, transparency in methodological detail, and further refinement and testing of published work and methodologies. The software should comply with accepted software development and imaging biomarker standards.
(2) Models and features should be tested to determine added prognostic and predictive accuracy compared to accepted clinical factors (e.g. tumour volume [23]). This safeguard is best done with clinical engagement.
(3) Multicollinearity of features should be investigated during model development with a training dataset. This will increase data variance description and improve precision of estimated regression coefficients, depending on modelling technique.
(4) Features should be tested for underlying dependencies using statistical analysis or by perturbing the data in controlled ways (as was demonstrated in this work). Perturbing data is a straightforward approach, but caution is warranted since it is difficult to control for confounding factors.
(5) Pre-processing of data should be done to ensure understanding of image signal quality (e.g. presence of dental artifacts). This will reduce the possibility of calculating erroneous features as demonstrated in the voxel-randomization portions of this work; and
(6) Manual contouring protocols should be included to describe any prevalent imaging signals (e.g. diagnostic MR images used in the determination of GTVs) utilized during delineation; particularly when volume remains highly influential in radiomic features.

Possible differences in radiomic feature implementation for the analysis of the four-feature prognostic model prevented exact reproduction of previous results, as the model itself is sensitive to small variations. However, signature analysis was permitted using PyRadiomics extracted Lung1 features to fit an updated model, MW2018, with an external-validation accuracy of 0.64 (0.60–0.68). Volume dominance in MW2018 was tested with data perturbation. Permuting image voxels decimated intensity feature components, while preserving volume and shape information; therefore, if MW2018 remained prognostic it was not because of patient specific intensity values and texture present in the GTV, but because of the GTV. In this scenario, MW2018 remained prognostic with c-index = 0.60 and 0.64 for Lung1 and H&N1, respectively, with standard deviations less than 0.006 across different voxel-randomization iterations. This implies that image intensity dependent features (energy, GLNU and HLH-GLNU), in conjunction with the cox model, did not capture relevant tumour phenotype information associated with texture, but were capturing the clinician delineated tumour volume, an established prognostic factor [23]. Additionally, the volume feature alone was as prognostic as MW2018 in Lung1 and H&N1 (c-index = 0.60 (0.56–0.64) and 0.64 (0.59–0.68)). These findings demonstrate the importance of careful and controlled experimentation, and a deeper exploration of proposed features. It should be noted that this does not mean that these features cannot be prognostic under a different signature or method of measurement.

It was determined that volume was driving the prognostic accuracy of univariate energy, GLNU and HLH-GLNU when the features were normalized for volume and had decreased univariate prognostic accuracies (Table 1 and Supplementary Figs. 3 and 4). This is consistent with recent published work demonstrating strong correlation between tumour volume and radiomic features in H&N [6] and lung [16] cancer. However, it should be noted that only linear dependencies were explored and higher order

dependencies may affect other features. This dependency may be the result of the initial feature selection process which selected features that remained stable between two scans [7]; therefore, biasing selection towards volume dependent features that would otherwise be unstable, such as wavelet features. Furthermore, high VIF values for GLNU and HLH-GLNU in the Lung1 dataset indicate that they describe the same data variance. Future model development should investigate structural multicollinearity amongst features to yield more stable models, while also potentially increasing prognostic accuracy above univariate tumour volume by including features that provide additional information. However, caution is required to ensure that unintended analysis of validation data does not occur, particularly when performing internal validation in the form of hold-out sets and out-of-bag bootstrapping.

The volume dependence of the signature's intensity based features may also be attributed to a variety of experimental processes, namely discretization of images and the original feature selection method. For this work we used a fixed bin width of 25 for discretization, the same methods employed by Aerts et. al. [7]. Discretization of images is used for texture analysis and image noise reduction [14], and requires appropriate bin widths to avoid over or under binning an image. A bin width of 25 is appropriate for the original images, but dynamic binning is the preferred method for wavelet images since the HU scale of the image is lost after filtering. Using narrow bin widths the images are subjected to over binning, resulting in the HLH-GLNU feature essentially counting the number of voxels in the ROI. Furthermore, interpolation to uniform voxel sizes are important in wavelet filtered images to ensure spatial frequencies are the same in every direction (although care must be taken that new artifacts are not introduced [24,25]), a method that was not used in our, or Aerts et. al's [7], work. Methods for voxel size interpolation are readily available in platforms like PyRadiomics, but dynamic binning methods are still needed; until these methods are standardized and implemented in radiomic platforms, utilization of our proposed methods will aid in safeguarding signatures and models from features with underlying volume dependencies.

Furthermore, manual tumour contours need to be considered since they are often determined from more than signal delineation on a single modality. Best practices for clinical contouring involve embedded and confounding parameters that influence delineations in ways that are hard to quantify, including additional patient characteristics, patterns of spread, and disease atlases; this is particularly true for H&N cancer [26,27]. Utilization of semi and fully-automated contouring software [28] could provide a more objective disease delineation, but could not be combined with clinical contours since they would not contain the nuanced information that is often important in clinical contours. Therefore, involvement of clinicians and detailed descriptions of the contouring protocols employed in radiomic analyses would ensure the importance of various features is appropriately attributed.

This work demonstrates the importance of being cautiously optimistic about radiomic signatures. Our findings do not imply that texture is not prognostic or predictive, only that enforced usage of standardization protocols and the adherence to the above safeguards is essential in avoiding incorrect feature interpretation and overly optimistic claims of robustness and generalizability. Radiomics is a relatively young field of study and requires transparency through high fidelity sharing of data and software to ensure progressive science. Radiomics dramatically expands our capacity to integrate imaging information into personalized cancer medicine, but it also represents a methodological shift within cancer research towards employment of highly automated approaches. The establishment of safeguards will serve to accelerate true progress in radiomics by educating researchers on the pitfalls and subtleties of these approaches and by encouraging broader engagement in the development and evaluation of the results.

Acknowledgements

The authors thank Dr. John Waldron and Mr. Biu Chan for their assistance in obtaining and curating the head and neck dataset used in this study. This work was supported by the Natural Sciences and Engineering Research Council to MLW; the Strategic Training in Transdisciplinary Radiation Science for the 21st Century program to MLW; the Canadian Institutes for Health Research; the Ontario Institute for Cancer Research; and the Terry Fox Research Institute. DAJ is the Mary and Orey Fidani Family Chair in Radiation Physics.

Conflict of interest statement

Dr. Dekker has a patent pending for systems, method and devices for analysing quantitative information obtained from radiological images. All other authors have no disclosures related to this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2018.10.027.

References

[1] O'Connor JPB, Rose CJ, Waterton JC, Carano RAD, Parker GJM, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. Clin Cancer Res 2015;21:249–57.
[2] Orel SG, Kay N, Reynolds C, Sullivan DC. BI-RADS categorization as a predictor of malignancy. Radiology 1999;211:845–50.
[3] Chong Y et al. Quantitative CT variables enabling response prediction in neoadjuvant therapy with EGFR-TKIs: Are they different from those in neoadjuvant concurrent chemoradiotherapy? PLoS One 2014;9:1–8.
[4] Fried DV et al. Prognostic Value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. Radiat Oncol Biol 2014;90:834–42.
[5] Coroller, TP, et al., Radiomic phenotype features predict pathological response in non-small cell lung cancer, vol. 119, pp. 480–486, 2016.
[6] Vallières, M, et al., Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer, pp. 1–33, 2017.
[7] Aerts HJWL et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
[8] Grossmann P et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017:1–22.
[9] Panth KM, et al., Is there a causal relationship between genetic changes and radiomics-based image features ? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells, vol. 116, pp. 462–466, 2015.
[10] Leijenaar RTH et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. Acta Oncol (Madr) 2015;54:2016.
[11] "MATLAB." The MathWorks Inc., Natick, USA; 2013.
[12] Bogowicz M et al. Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models. Radiother Oncol 2017.
[13] van Griethuysen, JJM, Al, E, Computational radiomics system to decode the radiographic phenotype," Submitted.
[14] Aerts HJWL, et al., Data From NSCLC-Radiomics. The Cancer Imaging Archive.".
[15] Zwanenburg, A, Leger, S, Vallieres, M, Lock, S, Image biomarker standardisation initiative, arXiv161207003; 2016.
[16] Fave X et al. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Transl Cancer Res 2016;5:349–63.
[17] Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: An R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics 2011;27:3206–8.
[18] Penciana MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. Stat Med 2004;23:2109–23.
[19] Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. Stat Med 2015;34:685–703.

[20] Kutner MH, Machtsheim CJ. 10.5 multicollinearity diagnostics – variance inflation factor. Applied linear regression models. McGraw-Hill Irwin; 2004.
[21] Farrar DE, Glauber RR. Multicollinearity in regression analysis: the problem revisited. Rev Econom Stat 1967;49:92–107.
[22] Leger S et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci Rep 2017;7:13206.
[23] Chao KSC et al. Intensity-modulated radiation therapy for oropharyngeal carcinoma: Impact of tumor volume. Int J Radiat Oncol Biol Phys 2004;59:43–50.
[24] Aganj I, Yeo BTT, Sabuncu MR, Fischl B. On removing interpolation and resampling artifacts in rigid image registration. IEEE Trans Image Process 2013;22:816–27.
[25] Tsao J. Interpolation artifacts in multimodal image registration based on maximization of mutual information. IEEE Trans Med Imaging 2003;22:854–64.
[26] Grégoire, V, et al., Delineation of the neck node levels for head and neck tumors: a 2013 update, vol. 110, pp. 172–181, 2013.
[27] Grégoire V, Eisbruch A, Hamoir M, Levendag P. Proposal for the delineation of the nodal CTV in the node-positive and the post-operative neck. Radiother Oncol 2006;79:15–20.
[28] Velazquez ER et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. Sci Rep 2013;3:3529.