Physica Medica 71 (2020) 24–30

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Machine learning helps identifying volume-confounding effects in radiomics

Alberto Traverso[a,b,*,1], Michal Kazmierski[a,*], Ivan Zhovannik[a,c], Mattea Welch[a,b], Leonard Wee[a], David Jaffray[b], Andre Dekker[a], Andrew Hope[b]

[a] Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
[b] Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto, Canada
[c] Department of Radiation Oncology, Radboudumc, Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Radiomics
Machine learning
Predictions
Lung
Head and neck

ABSTRACT

Purpose: Highlighting the risk of biases in radiomics-based models will help improve their quality and increase usage as decision support systems in the clinic. In this study we use machine learning-based methods to identify the presence of volume-confounding effects in radiomics features.

Methods

841 radiomics features were extracted from two retrospective publicly available datasets of lung and head neck cancers using open source software. Unsupervised hierarchical clustering and principal component analysis (PCA) identified relations between radiomics and clinical outcomes (overall survival). Bootstrapping techniques with logistic regression verified features' prognostic power and robustness.

Results

Over 80% of the features had large pairwise correlations. Nearly 30% of the features presented strong correlations with tumor volume. Using volume-independent features for clustering and PCA did not allow risk stratification of patients. Clinical predictors outperformed radiomics features in bootstrapping and logistic regression.

Conclusions

The adoption of safeguards in radiomics is imperative to improve the quality of radiomics studies. We proposed machine learning (ML) – based methods for robust radiomics signatures development.

1. Introduction

Radiomics, the automated extraction of quantitative descriptors from medical images, has demonstrated promising prognostic and predictive results for overall survival [1], distant metastases [4] and cancer biology [6]. After an initial phase of enthusiasm related to the introduction of this technology in the medical domain, investigation of the weakness and drawbacks of the new methodology always follows. These discussions are constructive and represent part of the scientific process to mature a technology, especially if it is meant to be clinically applicable.

In the radiomics scenario, recent publications warned about the presence of biases and potential risks that could be associated with radiomics-based models. In Chalkidou et al. [3], the authors pointed out that the usage of an elevated number of features combined with arbitrary feature selection cut-offs, might produce the undesired problem of multicollinearity, which leads to model over-fitting, often related to false discovery rates. The problem is that all radiomics computational packages compute hundreds to thousands of radiomics features, which often do not differ in their definitions, but are the same formulas computed by perturbing the original image with digital filters. This hyperspace of correlated features is usually much larger than the outcomes of interest, leading to models that are prone to overfitting and exposed to false positive associations [8]. Moreover, some radiomics features embed in their definition hidden confounding factors, which drive their prognostic/predictive power, but it is not immediately understood by inspecting the mathematical definitions of the features. A recent paper showed the presence of a strong volume-confounding effect in some radiomics signatures based on texture or statistical features [14]. The authors showed the randomization of grey level values still produced radiomics features able to have strong predictive power. This paper was the first one to introduce the concept of "safeguards" in

* Corresponding authors.
E-mail address: alberto.traverso@maastro.nl (A. Traverso).
[1] These authors equally contributed to the manuscript.

https://doi.org/10.1016/j.ejmp.2020.02.010
Received 30 October 2019; Received in revised form 12 February 2020; Accepted 13 February 2020
Available online 20 February 2020

radiomics studies. Understanding and evaluating the correlations between radiomic features and clinical prognostic variables is fundamental to evaluate the added value of imaging features compared to the previously mentioned factors. In a recent study [7], the authors investigated the complementary nature of heterogeneity quantified by imaging features and tumour volume in FDG-PET from multi-site cancers. They showed that volume and imaging features were both independent prognostic factors for Non-small Cell Lung Cancers (NSCLC) for volumes above 10 cm(3), with complementary information increasing substantially for larger tumour volumes. However, when smaller volumes were considered as in oesophageal cancers, the complementary value was degraded because of the presence of smaller volumes. Again, another study in FDG-PET [2], but for cervical cancers investigated the effect of e small tumour volumes on studies of intertumoral heterogeneity of tracer uptake. The authors used a computer simulation to isolate the effects of tumour volume on the image local entropy. They concluded that inclusion of tumour volumes below 45 cm (3) can profoundly bias comparisons of intra-tumoral uptake heterogeneity metrics. From the cited studies, to fully exploit the complementary prognostic/predictive power of imaging features it is imperative to benchmark them with respect for example to tumour volume. In fact, additional prognostic factors should be added to an existing model, since the introduction of redundant information could be dangerously prone to overfitting. By taking the previous studies as support, in this paper we intent to provide the radiomics community with a machine-learning based framework to evaluate complimentary role of imaging features when benchmarking with other prognostic factors, such as for example tumour volume.

We investigated how machine learning techniques can be used to discover the presence of volume-confounded features, effectively applying radiomics safeguards.

Machine learning methods are often used in the form of supervised methods, where classifiers are trained to learn associations between radiomics features and outcomes (labels). Large efforts have been dedicated to tuning classifiers, but there is no guarantee that biases will be uncovered. On the contrary, unsupervised methods do not look at labels and only utilize the original radiomics features. These methods are very popular in genomics studies, but not often used in radiomics studies. In this work we show how a combination of unsupervised and supervised methods can be used to introduce safeguards to radiomics studies.

2. Methods

2.1. Datasets

We used two retrospective public data sets for the analysis:

1) Lung1: 421 NSCLC (Non-Small Cell Lung Cancer) patients treated with concurrent chemo-radiotherapy. Computed Tomography (CT) scans of the patients and manually delineated contours of the primary Gross Tumor Volume (GTV) in form of DICOM and RTSTRUCT files were available. The dataset is available for download at the XNAT repository (https://xnat.bmia.nl) and on the TCIA archive (http://doi.org/10.7937/K9/TCIA.20%0A15.L4FRET6Z). The dataset is the same used in Aerts et al. [1], Shi et al., [10].
2) HN1: 132 CT scans of oropharynx and larynx squamous cell carcinoma patients treated with concurrent chemo-radiotherapy and manually delineated contours of the primary gross tumor volume (GTV) in form of DICOM and RTSTRUCT files were available. The dataset is available for download at the XNAT repository (https://xnat.bmia.nl) and the TCIA (https://doi.org/10.7937/tcia.2019.8kap372n). The dataset is the same used in Aerts et al. [1].

Additionally, clinical variables including: TNM, AJCC staging information, age, sex, as well as overall survival (OS) with a 3-year follow up were available.

2.2. Radiomics features extraction

We used the open source software PyRadiomics v2.2.0 [15] to extract imaging features from each GTV. Pyrex (https://github.com/zhenweishi/Py-rex), an extension of PyRadiomics was used to handle DICOM RTSTRUCT files as input, by generating a binary segmentation mask from the contour data [9]. For LUNG1 and HN1 datasets we used extraction parameters suggested in Aerts et al. [1]. A detailed description of the computational settings is provided in the Supplementary material. To further evaluate the impact in the results of the aggregation method of texture features, we compared the default “3Daverage” with the other most commonly used “3Dmerging” method.

Following features classes were extracted: statistical first order (FO), shape metrics (SM), texture features (TA) including Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighbouring Gray Tone Difference Matrix (NGTDM), Gray Level Dependence Matrix (GLDM), wavelet features (WF) for all of the above features excluding shape, computed using all combinations of applying either a High or Low pass filter in each of the three dimensions. In total, we extracted 841 features from each image volume (18 FO, 13 SM, 23 GLCM, 16 GLRLM, 16 GLSZM, 14 GLDM, 5 NGTDM, and 736 WF). The shape feature volume of each GTV was approximated by multiplying the number of voxels in the region of interest (ROI) by the volume of one voxel.

2.3. Elimination of redundant features

Pairwise feature inter-dependencies were evaluated using the Spearman rank correlation coefficient (ρ) The ρ metric does not assume any a priori functional dependence for the data (contrary, for example, to the Pearson coefficient) and therefore it is able to catch complex functional dependencies between features. The redundant features (with $|\rho| \geq t$, where t is a chosen threshold value) were eliminated by randomly dropping one of the two features. Thresholds from 0 to 1 with a 0.05 increment step were used.

2.4. Cluster analysis

We used hierarchical clustering to discover groups of patients with similar radiomics signatures. The optimal number of clusters (k) was determined using the consensus clustering method [16]. Briefly, clustering is repeated multiple times for different values of k using random sub-samples of the data. The value of k resulting in the most stable clusters (i.e. least change in cluster assignment for each observation across samples) is selected. We compared the distributions of clinical variables and GTV volume between clusters. In addition, we computed the Kaplan-Meier estimator of overall survival in each cluster. The log-rank test is a standard procedure to assess the statistical significance of difference between survival function estimates (with p-values corrected for multiple comparisons, using the FDR (False discovery rate) correction method (Benjamini-Hochberg procedure) [16].

2.5. Principal component analysis (PCA)

Principal component analysis (PCA) is an unsupervised method aiming to discover the sources of variance in the data. PCA identifies the directions of largest variability in the original dataset (called principal components). PCA is useful to determine if there is a confounding factor intrinsically present in the computed features, which is driving the variance in the data. The principal components are linear combinations of features and are ordered by the amount of total variance they explain. Thus, the first principal component represents the predominant pattern in the data and its strength is captured by explained variance. PCA can be used to identify latent variables sources of variability not

observed directly but nevertheless captured by the features. For example, if a suspected confounding variable is highly 1correlated with the first principal component, it is likely to be the true source of variation. A more comprehensive overview of the technique can be found in Traverso et al. [12].

2.6. Feature selection and modelling

To investigate the predictive value of volume-independent ($\rho \leq 0.1$) imaging features, we used them in combination with clinical variables (age, T, N, M stages, AJCC stage and tumour volume) in a binary logistic regression. We applied the model to two-year overall survival prediction. To determine the relative importance of features, as well as the stability to perturbations in the input, we applied a bootstrap-based method as detailed in [17]. Briefly, the model is refit on multiple bootstrap re-samples of the data and the order in which a feature is important for a model is obtained using Recursive Feature Elimination (RFE). The importance of each feature, as well as correlations between features, can be identified easily by visualizing the resampling results. Furthermore, the overall importance of each feature can be identified by aggregating the results across bootstrap resamples. Bootstrap-based variable selection analysis increases the reliability of reported models. All the statistical analysis was performed in Python v3.7.5 using the statistical package scikit-learn v0.21.3. Statistical significance was set at $p < 0.05$.

The workflow is briefly summarized in Table 1 (further commented in the discussion section) and it is composed by the following sequential steps that were adopted in this study: 1) evaluation of pairwise correlations between tumour volume using Spearman rank analysis and drop highly correlated features; 2) use the reduced list to perform hierarchical clustering and evaluating distribution of clinical variables (or confounding factors) in the clusters. Use PCA to select components that explain the largest percentage of variance, but still evaluating correlations between components and confounding factors; 3) to address sample biases use bootstrap techniques with RFE (Recursive Feature Elimination) and force in the model the presence of clinical prognostic factors. Build the final signature by selecting the most selected features.

3. Results

3.1. Feature correlations

Pairwise Spearman correlation between features revealed a high level of inter-dependence in the NSCLC dataset, with over 80% correlating with at least one other feature at $|\rho| > 0.9$ (Fig. 1a). Furthermore, nearly 30% showed correlation with tumour volume greater than 0.75 (Fig. 1c). A similar correlation was observed in the HN1 dataset (Fig. 1b and 1d). [R1] SupplementaryTable S1 lists all the radiomic features that presented a Spearman correlation $|\rho| \geq 0.8$ with GTV.

3.2. Clustering and PCA analysis

Using all the feature set, the patients could be stratified into two groups with significantly different survival times (log-rank $P < 10–6$, Fig. 2a). The difference in tumour volume distribution between clusters was highly significant (permutation $P < .001$, Fig. 2a). Removing features moderately correlated with volume (with Spearman $|\rho| > 0.6$) still allowed for cluster separation by OS ($P < 10–6$, Fig. 2b); however, the volume difference between clusters remained highly significant ($P < .001$, Fig. 2b). Using only volume-independent features ($|\rho| < 0.1$) the groups could not be separated by survival ($P = .8$, Fig. 2c) or tumour volume ($P = .9$, Fig. 2c). Groups could be separated by survival [R2] using as only input feature the computed GTV (log-rank $P < 10–6$) with no statistically significant differences between the full signature and GTV signature as shown in Fig. 2d.

As per this experiment it is possible to appreciate a degradation of performances when slowly removing features that are highly correlated with tumour volume, finally reaching a point ($(|\rho| < 0.1$) where only volume independent features are left, but no stratification is possible.

The first principal component (PC) extracted from all feature signature correlated with volume (Spearman $\rho = 0.78$, Fig. 3a). The first 2 PCs explained over 50% of the total variance, reflecting the large number of volume-correlated features. The latent volume effect was still present when moderately ($|\rho| < 0.6$) correlated features were used (correlation with volume: $\rho = -0.37$ for PC 1 and $\rho = 0.79$ 140 for PC 2, Fig. 3b), explaining the significance between-cluster differences in tumour volume. Finally, there was no volume-dominant effect in features independent ($|\rho| < 0.1$) from volume ($\rho = 0.01$ for PC 1 and $\rho = 0.05$ for PC 2, 143 Fig. 3c). Due to a smaller number of cases in the HN1 dataset, only one cluster could be reliably identified. In PCA, we found a dominant volume effect in full signatures ($\rho = -0.91$, Fig. 3a right) similarly to the NSCLC dataset. Crucially, the effect was present even in moderately correlated features (correlation with volume: $\rho = -0.52$ for PC 1 and $\rho = -0.58$ 149 for PC 2, 3b right). Again, the effect was not present in non-correlated features 150 ($\rho = -0.05$ for PC 1 and $\rho = 0.03$ for PC 2, Fig. 3c right).

3.3. Feature selection and modelling

Fig. 4 shows the order of selection in each bootstrap dataset (1000 replications in total) alongside the frequency of each feature entering the model first. In Lung1, volume enters the model first in most resampling iterations (84%), followed by T stage (which carries partially overlapping, but not identical information) and M stage (both 10%). In HN1, it is worth noting that the number of volume-independent features is larger. The most frequently selected feature is the overall stage (44%), followed by N stage and tumour volume (17% and 15% respectively). This reflects the higher importance of nodal involvement in head & neck squamous carcinomas for OS [18]. Interestingly, one

Table1
Suggested workflow for radiomics signature developments that incorporates safeguards.

Step #	Steps description	Method used	Address biased
1 – Redundancy and confounding factors analysis	Evaluate pairwise correlations between features by and tumour volume using ρ coefficients. Select a cut off for ρ and keep only non-redundant features	Spearman correlation coefficients (ρ)	Redundancy Confounding factors
2 – Clustering and PCA analysis	a) Use the new list of features as input for unsupervised hierarchical clustering b) Evaluate distributions of clinical variables in the clusters c) Select features that show significant differences between the clusters and explain most of the variance in the data (PCA)	Hierarchical clustering PCA	Dimensionality Reduction
3 – Outcome modeling	a) Use reduced list of features to develop the model. Choose the preferred classifier (e.g. SVM or logistic regression) combined with RFE b) Bootstrap the original dataset at least 1000 times and derive the most selected features c) Build the final model with most selected features	Bootstrapping RFE (Recursive feature elimination)	Sample bias

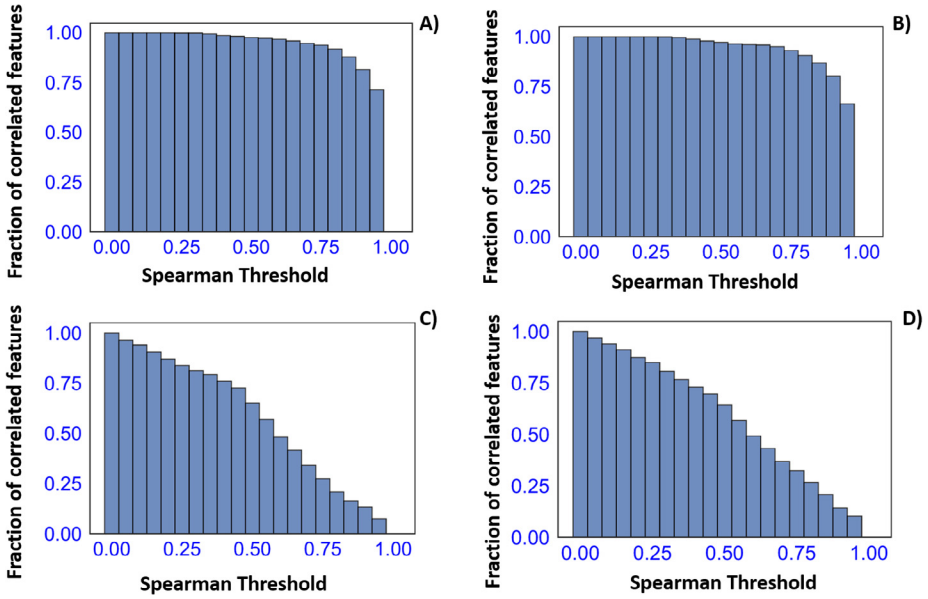

Fig. 1. Proportion of correlated features as a function of Spearman rank correlation threshold. (a) shows the proportion of features with pairwise correlation greater than the threshold value in the lung dataset. Percentage of features correlated with volume at a given threshold is shown in (c). The results were similar in HN1 (b, d).

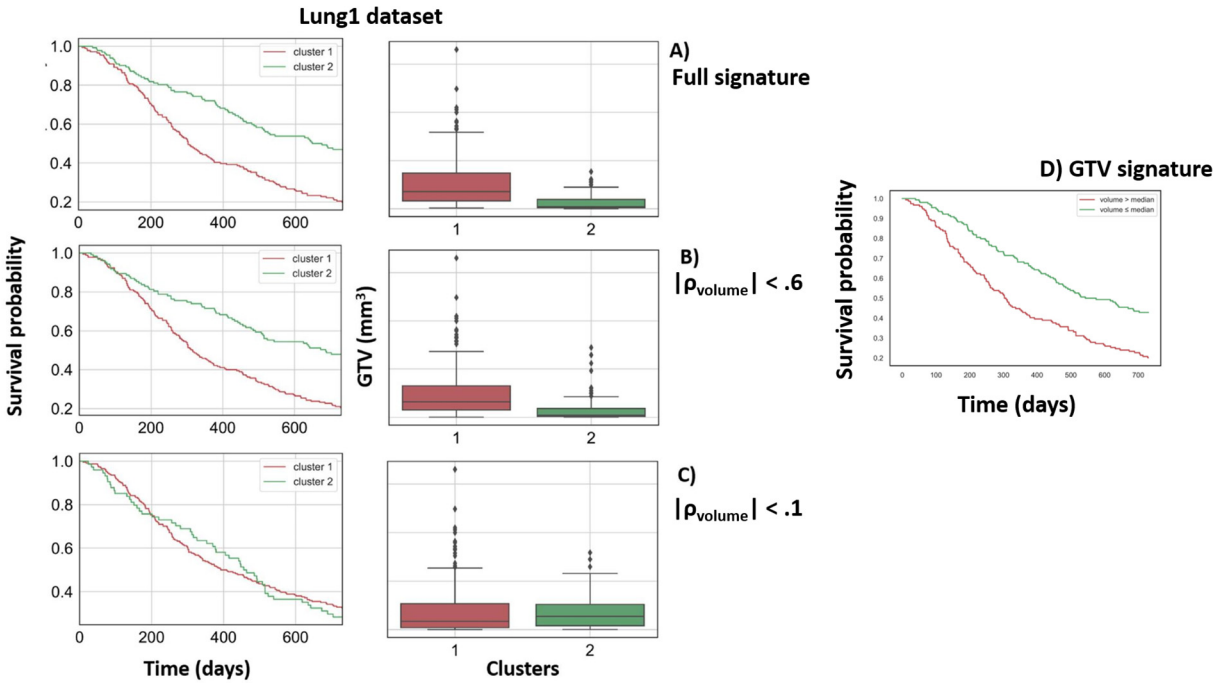

Fig. 2. Kaplan-Meier OS estimates and volume distributions for each cluster identified in the NSCLC dataset. (a) all original and wavelet features, (b) features moderately correlated with volume (defined as Spearman $|\rho| < 0.6$ with volume), (c) features not correlated with volume ($|\rho| < 0.1$), d) GTV signature.

imaging feature (GLDM-Gray Level Variance, correlation with volume 0.1) entered the model as first almost as frequently as volume (14%), indicating potentially complementary information. It is worth noting that our aim was not to create the optimal model, but rather to investigate the robustness of feature predictive performance. Results and conclusions remained unchanged when considering texture features computed with the "3Dmerging" aggregation approach.

4. Discussion

The evaluation of radiomics features multicollinearity and their benchmark with respect to accepted clinical prognostic factors is a needed safeguard. Our results show that radiomics features present strong inter-correlations, where texture features (TA) are usually more correlated between each other than first order (FO) features. Applying a wavelet filter augmented this problem, increasing therefore the dimensionality of problem to be solved and leading to a situation prone to overfitting.

Besides feature-feature correlations, a large percentage of radiomics features showed marked dependencies with tumour volume: 50% of total features had ρ volume > |0.6|, independently from the anatomical site considered (Fig. 1c and d). Again, TA features showed higher correlations with tumour volume than FO features. Three of highly correlated features were confirmed to be affected by strong volume correlations also in [14]. In HN1, texture features had slightly lower correlations with tumour volume than in Lung1.

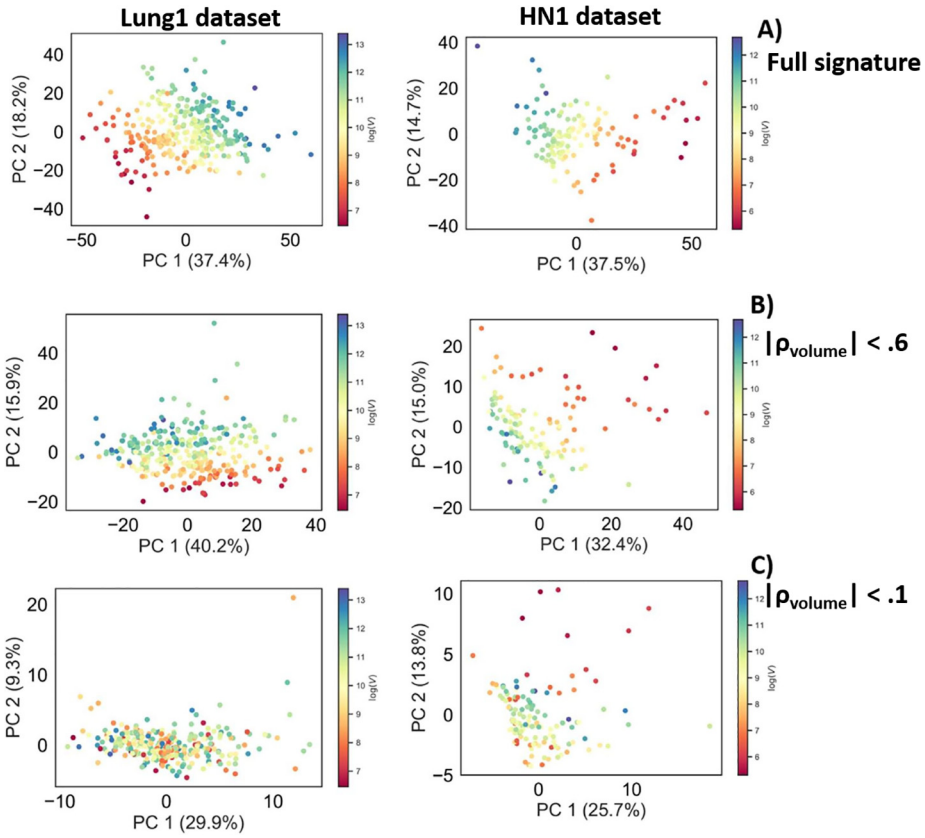

Fig. 3. Principal component analysis of (a) full feature signatures, (b) features moderately correlated with volume ($|\rho| < 0.6$), (c) volume-independent features ($|\rho| < 0.1$) in lung and head & neck datasets. The data is shown projected on the first 2 principal components and the proportion of variance explained by each component is indicated. Colors correspond to tumor volume. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Applying filtering decomposition of the original image did not eliminate the volume-effect. Our results confirmed that was no statistically significant difference of volume correlations between original and filtered features. Therefore, the usage of image filtering should carefully be adopted and justified to avoid an increase in the dimensionality of the features space to be reduced, without bringing any new information.

Tumour volume is a well-established and benchmarked prognostic factor for lung and head and neck cancers [5,11]. Therefore, correctly identifying if a feature or a combination of feature (e.g. signature) prognostic power is driven by a volume-confounded effect is fundamental to avoid spurious conclusions. With this evidence in mind we want again to clarify that the goal of our manuscript was not to discourage the community from building radiomics models and discard this effort since other predictors exist, but it was to provide this community with machine learning methods that could help achieving a good trade-off between explicability, transparency, parsimony, accuracy, and overfitting. Driver to reaching this trade-off, but still achieving good and robust performances is to identify the important explanatory variables.

Unfortunately, the majority of radiomics features come as complicated mathematical formulas, where identifying a direct and immediate dependence to tumour volume is far from trivial. In addition, two single features might not present strong dependencies to tumour volume, but their combination could.

In this study we showed how machine learning can be used to address the above-mentioned issue. Compared to the traditional radiomics workflow, we collocated machine learning at the top of the process, as a powerful instrument for exploratory analysis and acts as a safeguard against unanticipated cross-correlation with known prognostic features. The unsupervised methods of clustering and PCA present the following advantages: a) searching for patterns in the data without assuming any a-priori distribution or condition (i.e. without looking at the 'labels'); b) providing an intuitive way to retain pertinent information in the analysis and verify the main driver of it. When we cluster patients using all the radiomics features, the separation in terms of overall survival was statistically significant. However, the main reason of splitting can be attributed to strong volume differences between the clusters (Fig. 2). When we drop features correlated with tumour volume using a cut-off of 0.6, it was still possible to separate the two clusters in terms of OS, but with worse statistic. However, the clusters still had a predominant volume difference (Fig. 2). Finally, when considering only volume-independent features ($\rho \leq 0.1$) there was not significant splitting and no statistically significant difference between tumour volumes in the clusters. The results confirm that most of the radiomics features, when combined, led to spurious associations with tumour volume. Furthermore, volume-independent features alone did not allow stratification of patients into bad and good prognosis groups.

To further prove that the volume-latent effect is present independently from the unsupervised algorithm chosen, we repeated the PCA analysis but using the tSNE (t-distributed Stochastic Neighbouring Entities) [13]. It is another well known visualization method for high dimensional data, but compared to PCA, it uses a probabilistic approach. In our study, these two techniques were used as complimentary to further verify the found results. In fact, the same volume-latent effect was confirmed (figures available in the Supplementary material) also with tSNE.

Finally, we showed how bootstrap methods can be combined with supervised machine learning to evaluate feature significance. Furthermore, since bootstrap methods consider different subsamples of the original datasets, the risk of spurious associations, due to sample effects, is reduced. It is then possible to rank features according to their importance for the model by Recursive Feature Elimination. If a feature is important and has high prognostic value, it will often be selected, despite the chosen sample. A recent submitted publication to this journal related to radiomics-based model in head and neck cancers [14], showed that combining radiomics and clinical predictors did not lead to an elevate increase of performances. Similar results are found in

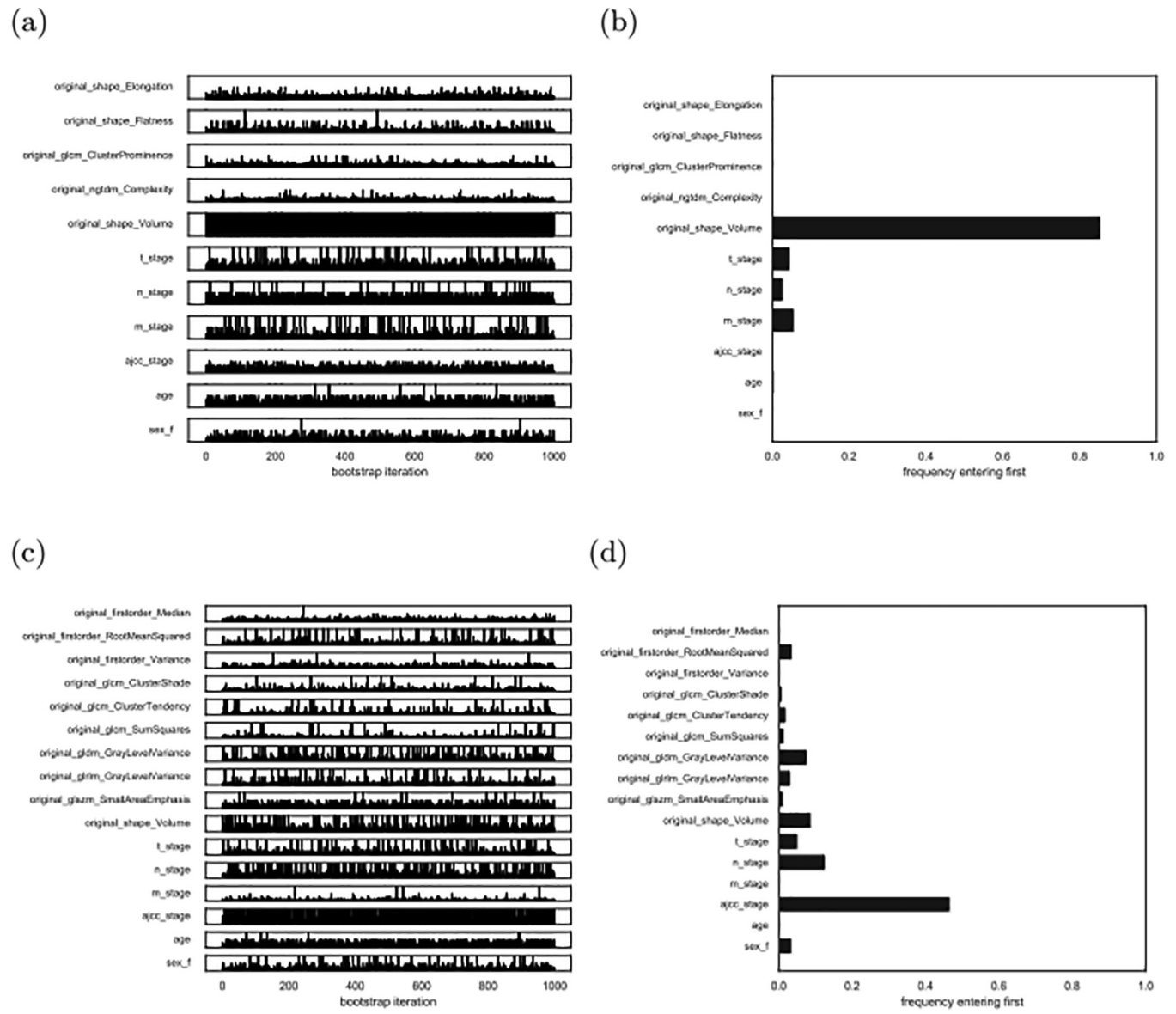

Fig. 4. Bootstrap-based evaluation of predictive power and stability of imaging and clinical features. (a) and (c) show the order of each feature entering the model across 1000 resampling iterations in Lung1 and HN1, respectively. The height of the bar is inversely proportional to the order of selection (therefore, filled bar indicates higher importance). (b) and (d) show the frequency of each feature entering the model first in both datasets.

our analysis also for the lung dataset: when considering only volume-independent features, tumour volume and t-stage outperformed each of the imaging features (Fig. 4a and b). In HN1 one radiomics feature was selected as often as other traditional clinical factors, but still the most frequent feature was nodal, showing that information outside the GTV (e.g. nodal involvement) plays a strong role in head and neck cancers. It is important to notice that it was out of this paper's scope to build the best model for predicting OS. Rather the aim was to provide the radiomics community with a method to benchmark radiomics predictors with accepted clinical factors and evaluate their stability with respect to a particular splitting of the datasets.

Finally, we provide safeguarding recommendations for signature developments in radiomics studies that build upon Welch et al. [14]: a) unsupervised learning methods (e.g. clustering and PCA) are preferable for exploratory analysis and dimensionality reduction with respect to traditional univariate and multivariate analysis; b) bootstrapping of radiomics predictors with accepted clinical factors provides a method to benchmark radiomics features and check the stability with respect to different sample sizes.

Table 1 summarizes a list of radiomics safeguards with suggestions of machine learning-based methodology for their applications.

[R3]While the results presented in this study remain valid only for the investigated clinical outcome (2-year OS), for the imaging modality (CT) and for the anatomical sites of lung and head and neck, the workflow presented in Table 1 can be extended as standard methodology for radiomic studies. We encourage the radiomic community to consider using unsupervised methods and the benchmarking of radiomic features with bootstrap techniques in their studies. Additional proven evidence of results found in this paper (e.g. degradation/contamination of prognostic power as a function of GTV/ feature dependencies) will help improving the quality of radiomic studies as well as re-thinking the definitions/role of some radiomic features.

Finally, it is worth mentioning some limitations of this study: a) due to limited availability we focused only on OS; b) the bootstrap modelling was limited only to logistic regression, but it could have extended also to other classifiers; c) the stated conclusions only apply to the studied anatomical sites (lung and head and neck) and for non CE (Contrast Enhanced) CT. The same conclusions might not be valid when

different modalities are considered (e.g. PET/CECT/MR) or applied to other anatomical sites, posing the urgent need to validate and share our methods with the radiomics community. Future works include addressing points a) and b) as well as considering volume-correction methods for improving signature developments in radiomics. We are planning to release the code as open source to incentive the community to adopt the presented methodology as benchmark for their studies.

5. Conclusion

In available datasets, volume confounds common radiomics analysis approaches. Volume, or other parameters which confound analysis, should be recognized during any radiomics workflow and dedicated safeguards should be built into analysis pipelines to identify and mitigate these risks.

[R3] Our study showed that by only using volume-independent features it was not possible to cluster patients in different survival groups.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2020.02.010.

References

[1] Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5. https://doi.org/10.1038/ncomms5006.

[2] Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. J Nucl Med 2014;55:37–42. https://doi.org/10.2967/jnumed.112.116715.

[3] Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One 2015;10:e0124165https://doi.org/10.1371/journal.pone.0124165.

[4] Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114:345–50. https://doi.org/10.1016/j.radonc.2015.02.015.

[5] Dejaco D, Steinbichler T, Schartinger VH, Fischer N, Anegg M, Dudas J, et al. Prognostic value of tumor volume in patients with head and neck squamous cell carcinoma treated with primary surgery. Head Neck 2018;40:728–39. https://doi.org/10.1002/hed.25040.

[6] Gardin I, Grégoire V, Gibon D, Kirisli H, Pasquier D, Thariat J, et al. Radiomics: principles and radiotherapy applications. Crit Rev Oncol/Hematol 2019;138:44–50. https://doi.org/10.1016/j.critrevonc.2019.03.015.

[7] Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med 2015;56:38–44. https://doi.org/10.2967/jnumed.114.144055.

[8] Matzner-Lober E, Suehs CM, Dohan A, Molinari N. Thoughts on entering correlated imaging variables into a multivariable model: application to radiomics and texture analysis. Diagn Intervent Imaging 2018;99:269–70. https://doi.org/10.1016/j.diii.2018.04.011.

[9] Shi Z, Traverso A, Soest J, Dekker A, Wee L. Technical note: ontology-guided radiomics analysis workflow (O-RAW). Med. Phys. 2019. https://doi.org/10.1002/mp.13844. mp.13844.

[10] Shi Z, Zhovannik I, Traverso A, Dankers FJWM, Deist TM, Kalendralis P, et al. Distributed radiomics as a signature validation study using the personal health train infrastructure. Sci Data 2019;6:218. https://doi.org/10.1038/s41597-019-0241-0.

[11] Takenaka T, Yamazaki K, Miura N, Mori R, Takeo S. The prognostic impact of tumor volume in patients with clinical stage IA non-small cell lung cancer. J Thorac Oncol 2016;11:1074–80. https://doi.org/10.1016/j.jtho.2016.02.005.

[12] Traverso A, Dankers FJWM, Osong B, Wee L, van Kuijk SMJ. Diving deeper into models. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of clinical data science Cham: Springer International Publishing; 2019. p. 121–33. https://doi.org/10.1007/978-3-319-99713-1_9.

[13] Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. Distill 1 2016. https://doi.org/10.23915/distill.00002.

[14] Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2018. https://doi.org/10.1016/j.radonc.2018.10.027.

[15] van Griethuysen JJM, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

[16] Vega-Pons S, Ruiz-Shulcloper J, et al. A survey of clustering ensemble algorithms. Int J Pattern Recogn Art Intel 2011;25(3):337–72. https://doi.org/10.1142/S0218001411008683.

[17] El Naqa I, et al. Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors. Int J Radiat Oncol Biol Phys 2006;15(64):1275–86. https://doi.org/10.1016/j.ijrobp.2005.11.022.