RESEARCH ARTICLE

Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC patients using FDG-PET/CT radiomics

Janna E. van Timmeren[1,2☯]*, Sara Carvalho[2☯], Ralph T. H. Leijenaar[1], Esther G. C. Troost[3,4,5,6,7,8,9,10], Wouter van Elmpt[2], Dirk de Ruysscher[2], Jean-Pierre Muratet[11], Fabrice Denis[11], Tanja Schimek-Jasch[12], Ursula Nestle[12], Arthur Jochems[1], Henry C. Woodruff[1], Cary Oberije[1], Philippe Lambin[1]

1 The D-Lab: Decision Support for Precision Medicine, GROW—School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands, **2** Department of Radiation Oncology (MAASTRO clinic), GROW—School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands, **3** OncoRay–National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Cal Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden–Rossendorf, Dresden, Germany, **4** Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, **5** Helmholtz-Zentrum Dresden–Rossendorf, Institute of Radiooncology—OncoRay, Dresden, Germany, **6** German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg, Germany, **7** National Center for Tumor Diseases (NCT) Partner Site Dresden, Dresden, Germany, **8** German Cancer Research Center (DKFZ), Heidelberg, Germany, **9** Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, **10** Helmholtz Association / Helmholtz-Zentrum Dresden–Rossendorf (HZDR), Dresden, Germany, **11** Centre Jean Bernard, Clinique Victor Hugo, Le Mans, France, **12** Department for Radiation Oncology, University Medical Center Freiburg, Freiburg, Germany

☯ These authors contributed equally to this work.
* j.vantimmeren@maastrichtuniversity.nl

OPEN ACCESS

Citation: van Timmeren JE, Carvalho S, Leijenaar RTH, Troost EGC, van Elmpt W, de Ruysscher D, et al. (2019) Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC patients using FDG-PET/CT radiomics. PLoS ONE 14(6): e0217536. https://doi.org/10.1371/journal.pone.0217536

Editor: Chin-Tu Chen, University of Chicago, UNITED STATES

Received: June 14, 2018

Accepted: May 11, 2019

Published: June 3, 2019

Data Availability Statement: Data originating from Maastro Clinic are available on request from the respective Institutional Review Board for researchers who meet the criteria for access to confidential data. Interested researchers may contact info@maastro.nl or Wouter van Elmpt at wouter.vanelmpt@maastro.nl.

Funding: Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n˚

Abstract

Background

Prognostic models based on individual patient characteristics can improve treatment decisions and outcome in the future. In many (radiomic) studies, small size and heterogeneity of datasets is a challenge that often limits performance and potential clinical applicability of these models. The current study is example of a retrospective multi-centric study with challenges and caveats. To highlight common issues and emphasize potential pitfalls, we aimed for an extensive analysis of these multi-center pre-treatment datasets, with an additional ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography/computed tomography (PET/CT) scan acquired during treatment.

Methods

The dataset consisted of 138 stage II-IV non-small cell lung cancer (NSCLC) patients from four different cohorts acquired from three different institutes. The differences between the cohorts were compared in terms of clinical characteristics and using the so-called ‘cohort

694812 - Hypoximmuno) and the QuIC-ConCePT project, which is partly funded by EFPI A companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. This research is also supported by the Dutch technology Foundation STW (n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL - n°673780), EUROSTARS (DART, DECIDE,), the European Program H2020-2015-17 (ImmunoSABR - n° 733008 and PREDICT - ITN - n° 766276), Interreg V-A Euregio Meuse-Rhine ("Euradiomics") and Kankeronderzoekfonds Limburg from the Health Foundation Limburg. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Author RL is CTO and shareholder of Oncoradiomics SA and co-inventor of a patent related to Radiomics. Author PL is member of the advisory board, shareholder of Oncoradiomics SA and co-inventor of two licenced & approved patents on Radiomics.

differences model' approach. Moreover, the potential prognostic performances for overall survival of radiomic features extracted from CT or FDG-PET, or relative or absolute differences between the scans at the two time points, were assessed using the LASSO regression method. Furthermore, the performances of five different classifiers were evaluated for all image sets.

Results

The individual cohorts substantially differed in terms of patient characteristics. Moreover, the cohort differences model indicated statistically significant differences between the cohorts. Neither LASSO nor any of the tested classifiers resulted in a clinical relevant prognostic model that could be validated on the available datasets.

Conclusion

The results imply that the study might have been influenced by a limited sample size, heterogeneous patient characteristics, and inconsistent imaging parameters. No prognostic performance of FDG-PET or CT based radiomics models can be reported. This study highlights the necessity of extensive evaluations of cohorts and of validation datasets, especially in retrospective multi-centric datasets.

Introduction

Prognostic models based on individual patient derived factors are essential to better estimate patient's outcome prior to or during treatment. These models help to improve individualized treatment decisions (personalized medicine) that may lead to better patient outcomes [1, 2]. Medical images are an example of an information source, which could be used to derive important patient-specific prognostic information. Large amounts of quantitative features can be calculated from medical images acquired during a patient's course of treatment, e.g. positron emission tomography (PET) or computed tomography (CT). This principle of extracting imaging features is called 'radiomics', which is a rapidly evolving field of interest [3–5]. Multiple studies have shown radiomics' potential to derive prognostic information for patient outcomes. Despite the promising results, radiomics faces multiple challenges [6]. An important challenge is the collection and acquisition of (large amounts of) suitable imaging data, which is difficult due to evolving technology, lack of standardization protocols and differences in cohorts and protocols between institutes.

Imaging data collected from a single institution often results in a more homogeneous dataset, e.g. the images are acquired using the same settings (with the institute's acquisition and reconstruction protocols) on the same scanner. Also, with regard to clinical characteristics the patient population is usually more homogeneous. However, a prognostic model developed based on these data might fail when applied to validation data from an external institute, due to the lack of transferability of the model (i.e. the model is specific for a population) [7, 8]. Therefore, one might argue for using more heterogeneous datasets to train the model to broaden the clinical applicability. In practice, it appears difficult to find these datasets and to validate such a model. Besides the variability in imaging and/or clinical characteristics within a dataset, the dimensionality of the data is often an issue, e.g. having over 1000 radiomic features with often only a limited number of patients. A review published in 2015 has shown that many published (radiomics) studies contain models based on a limited number of patients without validation on an independent dataset, resulting in high probabilities of false positive results [9].

In this study, PET and CT images were collected from non-small cell lung cancer (NSCLC) patients who underwent an ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography/computed tomography (PET/CT) scan prior to and during (chemo)radiotherapy to be able to evaluate radiomics' potential to assess early response. Previous studies have explored (early) response assessment using information derived from FDG PET/CT scans [10–15]. Since the PET/CT scans obtained during treatment are generally not acquired in clinical practice, it is difficult to gather large datasets of this type. Therefore, the data in the current study were collected from three different institutes where small sub-cohorts were available and is therefore a typical example of a multi-centric study with retrospectively collected data.

Retrospective studies, although collected from centers with high quality care, often suffer from heterogeneities within and between datasets, which is unavoidable in clinical practice. Therefore, the current study aims to evaluate the multicenter patient data in terms of cohort characteristics and prognostic performance, intending to highlight common issues and emphasizing pitfalls for future (radiomic) studies.

Materials and methods

Study cohort

The entire study cohort consists of 138 stage II-IV NSCLC patients, all treated with curatively intended (chemo)radiotherapy. The study population was divided into four different datasets from three different institutes. Datasets 1 and 2 originated from MAASTRO Clinic, Maastricht, the Netherlands. Dataset 3 is from Clinique Victor Hugo, Le Mans, France, and Dataset 4 was obtained from University Medical Center Freiburg, Freiburg, Germany. Dataset 1 was prospectively collected with approval of the Institutional Review Board of the Department of Radiation Oncology of Maastricht University Medical Center (Maastro Clinic) (clinicaltrials.gov NTC00522639). The original study that collected the data in Dataset 2 was approved by the Institutional Board of the Department of Radiation Oncology of Maastricht University Medical Center (Maastro Clinic), as stated in [12]. Dataset 3 was collected in a previous study [16] and approved by the appropriate Institutional Review Board. Dataset 4 was prospectively collected with approval of the Ethics Committee of the Albert-Ludwig University Freiburg, Germany Freiburg (clinicaltrials.gov NCT00697333).

All patients underwent an ^{18}F-FDG-PET/CT scan before radiotherapy for treatment planning purposes and an additional scan approximately two weeks after the start of treatment.

The scanning parameters are summarized in Table 1. Detailed cohort descriptions are provided in Supplementary Information: S1 Text.

For all patients, overall survival (OS), calculated from the start of radiotherapy, was collected to serve as a clinical endpoint in this study. A patient still alive at the time of study analysis was considered right-censored. Median follow-up times were calculated using the Kaplan-Meier estimator with reversed events, considering death as censored data.

Tumor segmentation and image analysis

The primary gross tumor volume (GTV) was delineated by experienced radiation oncologists on pre- and during-radiotherapy fused FDG-PET/CT images using the following WL settings, W:1700 L:-300 (lung) and for W:600 L:40 (mediastinum). These target delineations were not evaluated by other specialists and the reproducibility of these segmentations was not investigated in this study, as this segmentation process was just part of normal clinical practice.

Table 1. Scanning parameters. Scanning parameters for all scans included in the study.

Parameters		Dataset 1 (n = 100)	Dataset 2 (n = 62)	Dataset 3 (n = 54)	Dataset 4 (n = 60)
Manufacturer		Siemens	Siemens	Philips Healthcare	Philips Healthcare
CT	Tube voltage	120 kVp	120 kVp	120 kVp (n = 39)	120 kVp
				140 kVp (n = 15)	
	Tube current	120 mA (n = 3)	80 mA (n = 13)	Median [range]	Median [range]
		160 mA (n = 15)	336 mA (n = 49)	183 mA [115–277]	183 mA [35–337]
		173 mA (n = 1)			
		240 mA (n = 72)			
	Convolution kernel	B19f (n = 97)	B19f (n = 37)	B	B
		B30f (n = 2)	B30f (n = 12		
		B41f (n = 1)	B41f (n = 13)		
	Slice thickness	3 mm	3 mm	5 mm	1 mm (n = 16)
					2 mm (n = 16)
					3 mm (n = 4)
					4 mm (n = 24)
	Pixel spacing	0.98 mm	0.98 mm	1.2 mm (n = 44)	Median [range]
				1.4 mm (n = 10)	1.2 mm [0.7–1.4]
PET	Reconstruction algorithm	OSEM2D 4i8s	OSEM2D 4i8s	BLOB-OS-TF	BLOB-OS-TF
	Slice thickness	3 mm	3 mm	4 mm	2 mm (n = 38)
					4 mm (n = 22)
	Pixel spacing	4.07 mm	4.07 mm	4 mm	2 mm (n = 38)
					4 mm (n = 22)
	Injected FDG dose	Median [range]	Median [range]	Median [range]	Median [range]
		180 MBq [113–354]	302 MBq [175–482]	290 MBq [139–474]	340 MBq [198–434]
	Planned interval FDG injection–image acquisition	60 minutes	60 minutes	60 minutes	60 minutes

https://doi.org/10.1371/journal.pone.0217536.t001

FDG-PET/CT images were converted into standard uptake value (SUV) prior to analysis [17]. A total of 1295 radiomic features were extracted from CT images and 1400 radiomic features were extracted from PET images. Prior to radiomics feature extraction, all images were resampled into voxel dimensions 1×1×3 mm^3 for CT and 4×4×3 mm^3 for PET, which corresponded to the average voxel dimensions of all images rounded to the nearest integer. Resampling decreases the variability of radiomic features [18] and was performed using cubic interpolation. For one PET scan the resampling resulted in a segment splitting off from the original volume resulting in two separate volumes. As radiomic features have a different interpretation when extracted from multiple volumes, this patient was excluded from the analyses to avoid inconsistencies.

The image analysis included the investigation of different feature groups: 1) Morphological 'Shape', 2) Fractal, 3) Local Intensity 'LocInt', 4) Statistical features 'Stats', 5) Intensity-volume histogram 'IVH', 6) Textural features, including gray-level co-occurrence matrix 'GLCM', gray-level run length matrix 'GLRLM', gray-level size zone matrix 'GLSZM', neighborhood gray tone difference matrix 'NGTDM', gray-level distance zone matrix 'GLDZM', neighborhood gray-level dependence matrix 'NGLDM', 7) Laplacian of Gaussian filter 'LoG' (prior to group 4), and 8) Wavelet filter (prior to group 6). The feature descriptions of the feature groups (fractal, local intensity and intensity histograms) can be found in the Supplementary Information S2 Text. Other feature descriptions can be found elsewhere [19]. The IVH features were extracted from PET images only and these allow to retrieve the metabolic tumor volume (MTV), which is defined as the volume of voxels with an intensity above x% of

SUVmax within the lesion (from 10 to 90% ($MTV_{10\%}$—$MTV_{90\%}$)) [20] and total lesion glycolysis (TLG), by multiplying MTV by the corresponding mean SUV within the segmented volume: $TLG_{10\%}$—$TLG_{90\%}$ [20, 21].

Image analysis was performed in Matlab R2014a (The Mathworks, Natick, MA) using in-house developed software used for feature extraction. The absolute variation (abs) and percentage (rel) variation between subsequent scans were also derived, defined as:

$$\mathrm{Abs} = \mathrm{During}_{\mathrm{treatment}} - \mathrm{Pre}_{\mathrm{treatment}} \tag{1}$$

$$\mathrm{Rel} = \frac{\mathrm{During}_{\mathrm{treatment}} - \mathrm{Pre}_{\mathrm{treatment}}}{\mathrm{Pre}_{\mathrm{treatment}}} \tag{2}$$

In total, eight sets of radiomic features were derived from each dataset, being CT-scan1 (pre-treatment), CT-scan2 (during treatment), PET-scan1, PET-scan2, CT-abs, CT-rel, PET-abs and PET-rel.

Cohort comparison

To compare the four cohorts, two different approaches were applied. First of all, clinical characteristics were compared to test univariate cohort differences using a Wilcoxon rank test for continuous variables or a chi-square test for categorical variables. P-values below 0.05 were considered significant. Note that the p-values of these analyses were not corrected for multiple testing.

Secondly, the cohort differences (CD) model approach described in detail by van Soest *et al*. [7, 8] was used to assess the multivariate cohort differences by predicting to which cohort a patient belongs. It provides as summarizing measure of the level of generalizability of the model, ranging from reproducibility to transferability. Two CD-models were created. The first CD-model included two-year survival (binary variable) as an independent variable, as well as radiomic features of a model developed in the current study (see next section 'Model development'). The second CD-model also included the clinical variables 'Gender' and 'Overall Stage' as independent variables to investigate whether potential differences between the cohorts can be explained by those clinical parameters. As two-year survival was used as input variable, four patients of Dataset 2 were not included in this analysis due to shorter follow-up times. For the CD-model, stage was converted into dummy variables, using three categories II, IIIa or IIIb/IV. Stage II was used as the reference category.

The binary dependent variable of the CD-models was cohort A or B. A simple logistic regression was used to train the regression beta coefficients and predict to which cohort the data belongs. This procedure was applied to all possible combinations of two datasets. The performance of the CD-models for each combination of two datasets was evaluated by their Receiver Operating Curve (ROC). High CD Area under the Curve (AUC) values would indicate a large difference between the distributions in the cohorts and imply that the model tests transferability rather than reproducibility [7, 8].

Model development

Each of the four datasets was used to train a model, which was subsequently validated on the three remaining datasets. Moreover, a model was developed by combining all data into one large dataset (n = 138) which was split randomly into training (75%, n = 103) and validation (25%, n = 35), as proposed in [22].

For model development, a least absolute shrinkage and selection operator (LASSO) method was applied [23]. A 10-fold cross validation procedure was repeated 200 times to optimize the penalty coefficient lambda, i.e. to find the smallest error, and to stabilize the method. Inputs for LASSO were the total of extracted radiomic features from either CT-scan1, CT-scan2, PET-scan1, PET-scan2, CT-abs, CT-rel, PET-abs or PET-rel. The performance of the penalized Cox model was evaluated using Harrell's concordance index (c-index), for which 1 indicates perfect discrimination and a value of 0.5 no discrimination (no greater than chance expectation) [24]. Moreover, prognostic index (PI) values, defined as $\sum_i \beta_i x_i$, were calculated for all four datasets. These were analyzed to give insight in discriminative ability of the model, as proposed by Royston and Altman [25].

Classification comparison

For the combined dataset, five different classifiers were investigated according to the methodology described by Deist, Dankers *et al.* [26]. These include glmnet (penalized generalized linear models), rf (random forest), svm (support vector machine), LogitBoost (boosting) and rpart (regression trees). The neural network classifier was not investigated, as these required too long computation times for a large number of input parameters and extensive tuning. The classifiers deal with binary outcomes, therefore 'two-year survival' was used as endpoint. Since four patients in Dataset 4 did not reach the minimum of two years of follow-up, these patients were excluded for this specific analysis. The maximum number of repetitions was increased to 47 (limited by available calculation time on the computer) for this study and default tuning was turned off. All other settings were kept to the standards as described in [26].

All statistical methods were performed in R (version 3.4.3), using the packages *survcomp*, *survival* and *glmnet*. For the classification comparison, additional packages were used as described in [26]. P-values below 0.05 were considered significant.

Univariable analysis

The performance of commonly assessed PET metrics as potential prognostic factors was also investigated in an attempt to validate previous findings [10–16]. Also, radiomic features can be hard to interpret, whereas these PET metrics are well-known. A univariable Cox proportional hazard regression was computed for the percentage variation of most commonly assessed PET metrics: volume, maximum SUV (maximum image intensity value), mean SUV, peak SUV (maximum average SUV in a 1 cm^3 spherical volume), MTV$_{50\%}$ (volume above 50% of intensity) and TLG$_{50\%}$ (TLG for the volume above 50% of intensity). This analysis was performed on the combined dataset of 138 patients.

Results

Characteristics of the cohorts are summarized in Table 2, which shows that all clinical parameters were significantly different between one or more datasets, except age (indicated with the bold numbers). All patients from Datasets 1, 3 and 4 received concurrent chemoradiotherapy and no other treatment between the first PET scan and the start of radiotherapy (RT). In Dataset 2, 55% of patients received sequential chemoradiotherapy and one patient did not receive any chemotherapy. Because of the limited sample size available to us for this study, we have decided to not exclude outliers (e.g. PET scans separated by long time intervals) in an effort to make the dataset more homogeneous.

Median [range] survival was 1.6 [0.1–4.5], 2.3 [0.2–7.0], 1.8 [0.3–5.9] and 3.1 [0.1–5.2] years for Dataset 1, 2, 3 and 4, respectively (Fig 1). Median follow-up was 3.8 [2.5–4.5], 6.9 [6.6–7.0], 4.0 [2.6–5.9] and 3.4 [1.3–5.2] years for Dataset 1, 2, 3 and 4, respectively.

Fig 2 shows the ROCs of the CD-model for each combination of two datasets. AUC values ranged from 0.66 (Dataset 3 versus 4) to 0.89 (Dataset 2 versus 4). None of the confidence

Table 2. Patient characteristics. Patient characteristics for all four datasets used in this study. The bold numbers 1, 2, 3 or 4 indicate the datasets from which the variable was significantly different.

	Dataset 1 (n = 50)	Dataset 2 (n = 31)	Dataset 3 (n = 27)	Dataset 4 (n = 30)
Age [years]	-	-	-	-
Range (median)	35–86 (63)	46–82 (64)	41–76 (62)	47–83 (64)
Mean ± SD	62 ± 11	65 ± 9	61 ± 8	65 ± 9
Gender	**3**	**3**	**1,2,4**	**3**
Male	25 (50%)	22 (71%)	25 (93%)	19 (63%)
Female	25 (50%)	9 (29%)	2 (7%)	11 (37%)
Stage	**2,3,4**	**1**	**1**	**1**
II	-	2 (6%)	4 (15%)	1 (3%)
IIIa	17 (34%)	14 (45%)	16[a] (59%)	16 (53%)
IIIb	27 (54%)	15 (48%)	7 (26%)	13 (43%)
IV	6 (12%)	-	-	-
Histology	**3,4**	**3,4**	**1,2**	**1,2**
Adenocarcinoma	18 (36%)	6 (19%)	11 (41%)	9 (30%)
Squamous cell carcinoma	14 (28%)	9 (29%)	14 (52%)	18 (60%)
NSCLC Otherwise Specified	18 (36%)	16 (52%)	2 (7%)	3 (10%)
Radiotherapy [dose]	**2,3**	**1,3,4**	**1,2**	**2**
Range (median)	45–69 (69)	46–70 (61)	66–70 (66)	30–74 (66)
Mean ± SD	64 ± 6	61 ± 7	68 ± 2	66 ± 8
Chemotherapy	**2**	**1,3,4**	**2**	**2**
Concurrent	50 (100%)	13 (42%)	27 (100%)	30 (100%)
Sequential	-	17 (55%)	-	-
No	-	1 (3%)	-	-
Interval pre-PET–First RT [days]	**3,4**	**3,4**	**1,2,4**	**1,2,3**
Range (median)	4–16 (7)	2–13 (7)	5–93 (33)	2–37 (16)
Mean ± SD	7 ± 2	8 ± 2	38 ± 21	17 ± 7
Interval First RT–during-PET [days]	**2,3**	**1,3,4**	**1,2,4**	**2,3**
Range (median)	5–20 (15)	6–19 (8)	15–32 (21)[b]	14–24 (15)
Mean ± SD	15 ± 2	9 ± 3	21 ± 4	16 ± 3
Interval between PET scans [days]	**2,3,4**	**1,3,4**	**1,2,4**	**1,2,3**
Range (median)	19–27 (22)	10–24 (16)	21–110 (52)	22–59 (33)
Mean ± SD	22 ± 2	17 ± 3	59 ± 21	34 ± 8

[a] Includes one TxN2M0 patient, for which the merged structure between node and tumor was analyzed.

[b] Two patients' first PET scans were acquired more than 30 days after start of radiotherapy.

https://doi.org/10.1371/journal.pone.0217536.t002

intervals included 0.5, meaning that all AUC values were significantly different from 0.5 (not shown). The ROC curves of the CD-model, which included gender and overall stage, resulted in slightly higher AUC values, ranging from 0.77 to 0.92 (S1 Fig).

To assess prognostic performance, models were developed using all four different datasets once as training and the remainders as validation. The first models were trained on the image sets of the largest dataset, Dataset 1 (n = 50). Using LASSO, prognostic predictors were identified for the image sets CT-scan1, CT-rel and PET-abs. For the other image sets, no prognostic models were identified. Since Dataset 1 was the largest dataset and CT scans prior to treatment are currently most commonly used to assess prognostic performance, the model belonging to CT-scan1 of Dataset 1 was chosen to be used to construct the CD-model (Fig 2). This model, built using LASSO, consisted of 3 variables: 1) Wavelet LLH (Low Low High) Fractal sd

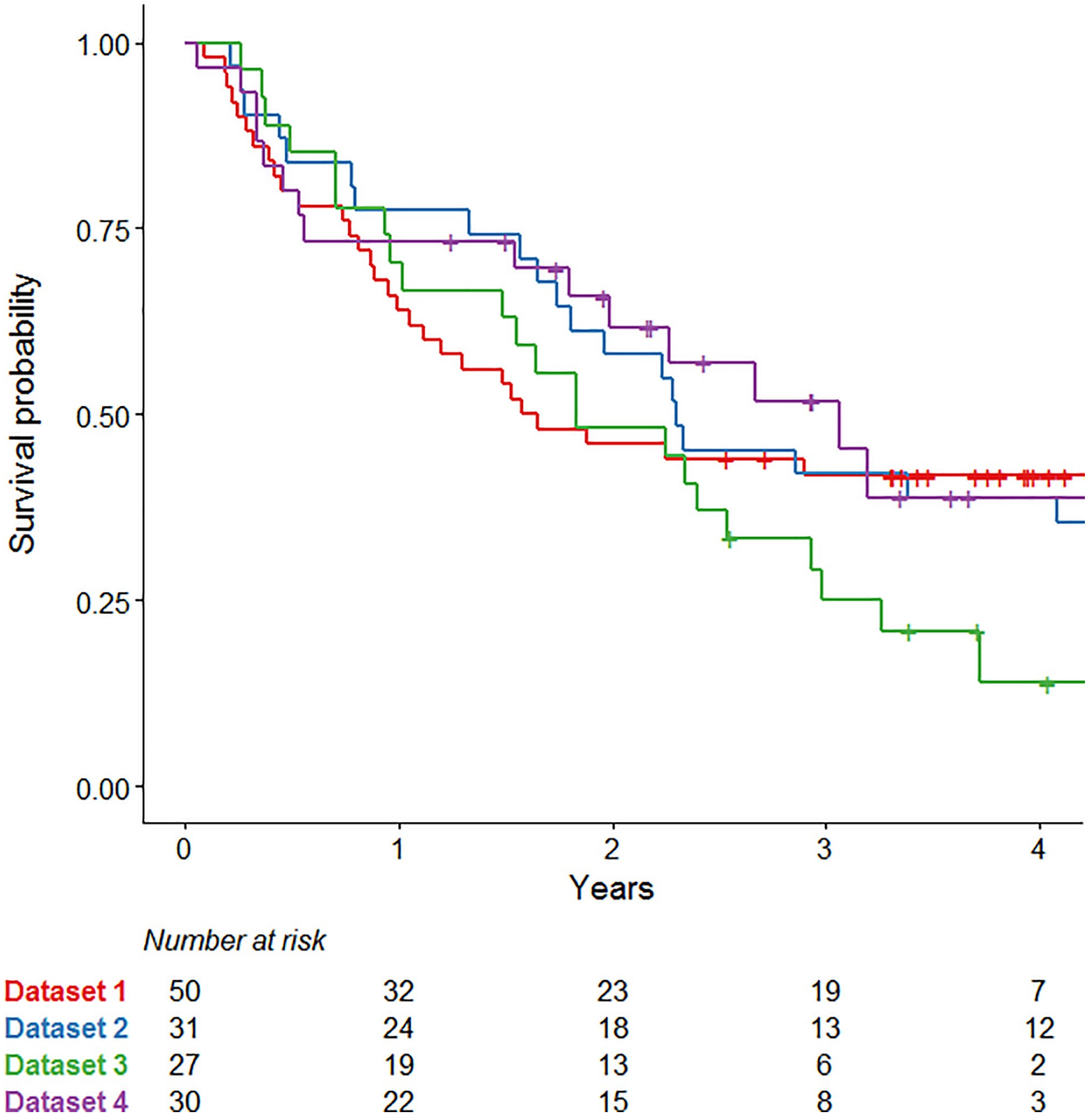

Fig 1. Kaplan-Meier curves. Kaplan-Meier curves for overall survival of all datasets.

https://doi.org/10.1371/journal.pone.0217536.g001

(standard deviation), 2) Wavelet LLH GLDZM IV (Inverse Variance) and 3) Wavelet LLH GLSZM IV. The corresponding beta coefficients of the Cox model were 1.38, $5.0^{*}10^{-17}$ and $2.9^{*}10^{-32}$, respectively. Fig 3 shows the ranges of the prognostic index (PI) for all four datasets for this model.

In Table 3 shows the results for Harrell's concordance index for the imaging features extracted from all image sets and the calculated differences. The performances of models trained on Dataset 2, 3 or 4 are shown in Supplementary Information S3 Text.

Furthermore, all data was combined into one large cohort and randomly split into training (n = 103) and validation (n = 35). For CT-rel and CT-abs, models were selected with c-indices

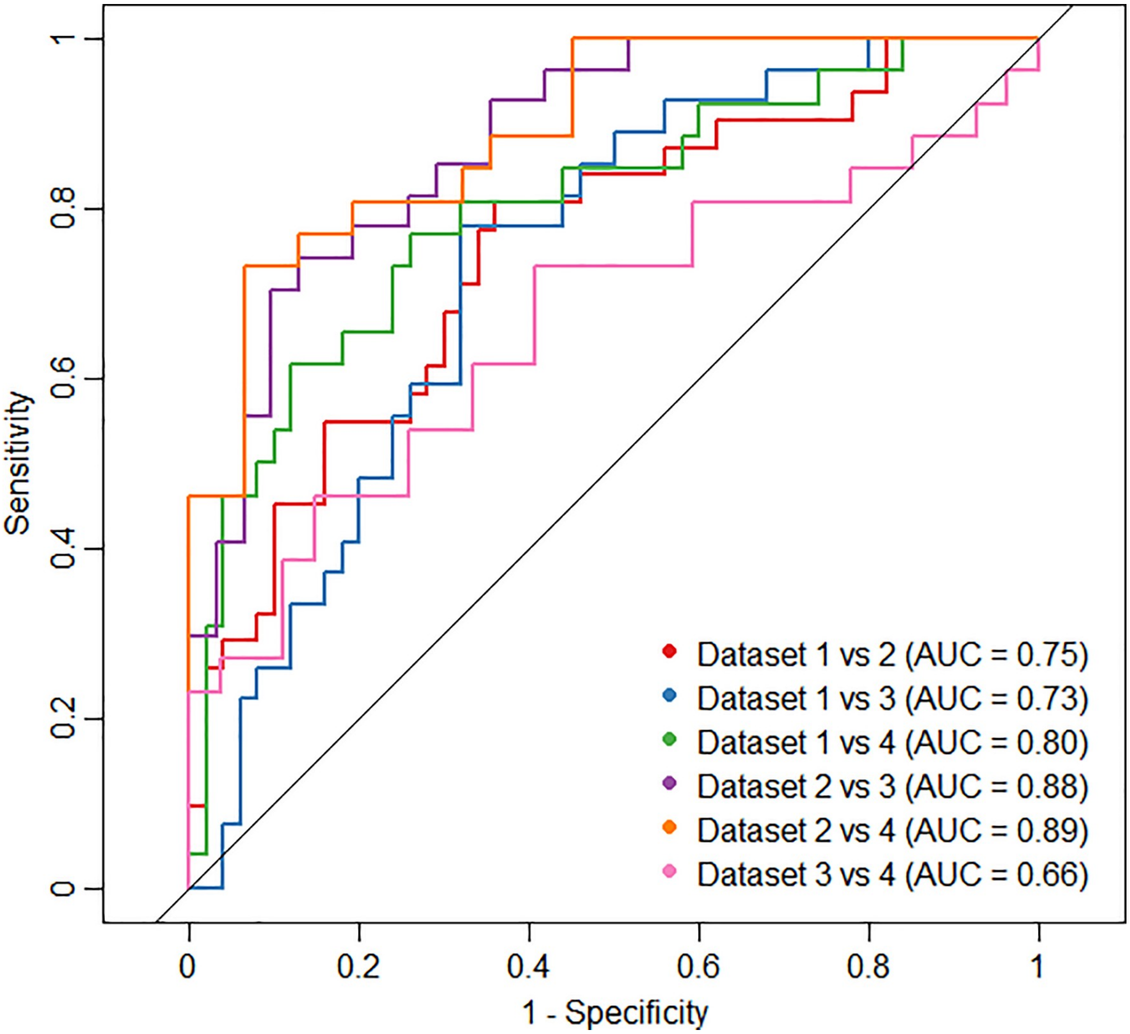

Fig 2. ROCs cohort difference model. Receiver Operator Curves (ROC) for the cohort difference (CD) model for each combination of datasets, including three radiomic features and 'two-year survival' as independent variables.

https://doi.org/10.1371/journal.pone.0217536.g002

significantly different from 0.5: 0.68 [95% C.I. 0.61–0.74] and 0.68 [95% C.I. 0.61–0.74], respectively. However, none of these models could be validated on the validation subgroup, reaching c-indices of 0.54 [95% C.I. 0.42–0.67] and 0.50 [95% C.I. 0.37–0.62]. For the other image sets, subsets of regressors were identified which resulted in predictions (i.e. relative risks) equal to 1 for all patients, meaning that the model did not contain any prognostic information.

Besides LASSO, different classifiers were investigated to further examine the prognostic value of the radiomic features for 2-year survival. The mean AUC values are shown in the heat-map of Fig 4.

Table 4 shows the results of the univariable analysis of the prognostic value of percentage variation of common PET imaging descriptors on the combined cohort of 138 patients. The results for hazard ratio (HR), corresponding confidence interval (CI) and the c-index were shown.

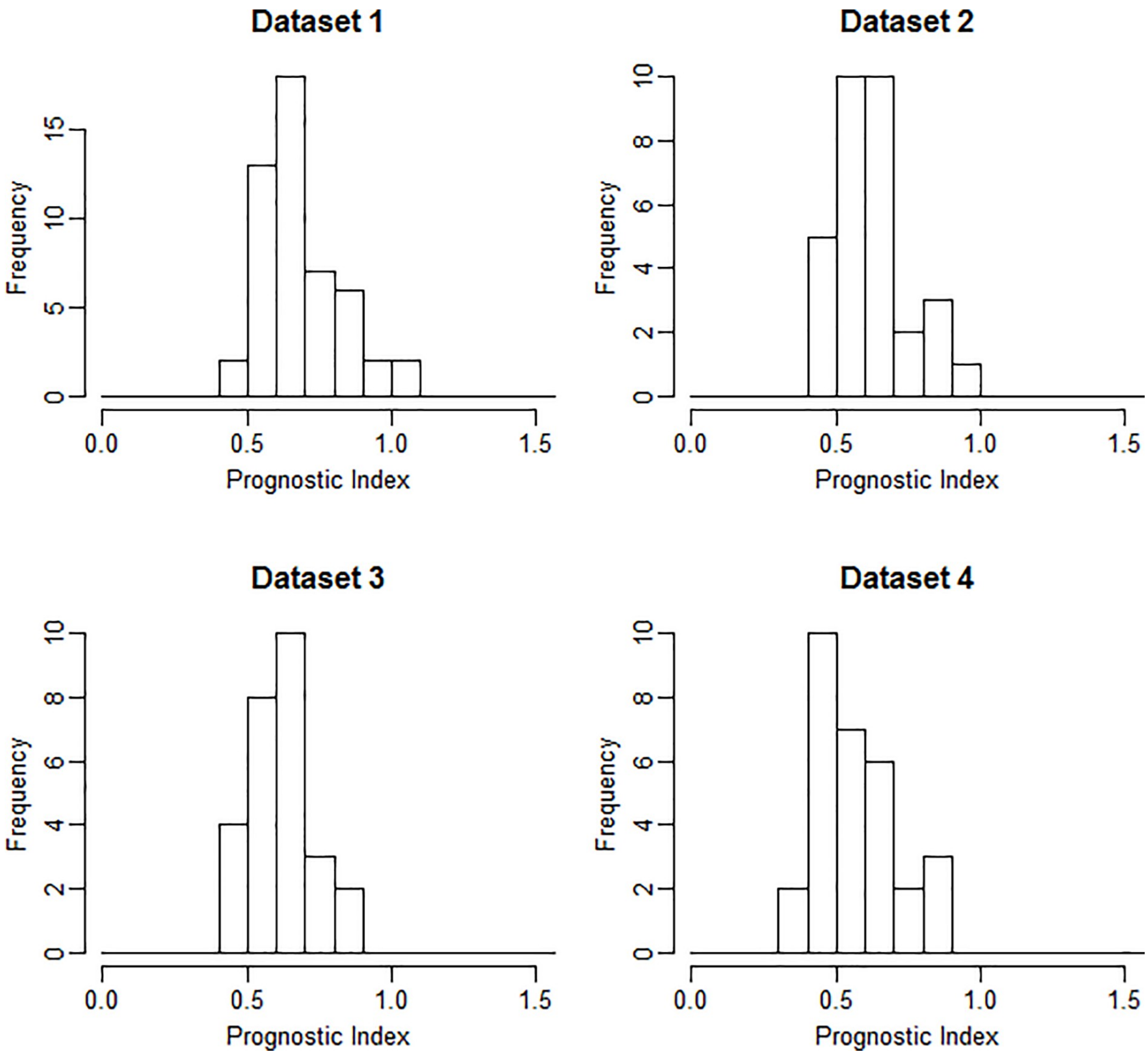

Fig 3. Prognostic Index. Prognostic index (PI) ranges for all datasets based on the model developed on CT–scan1, using Dataset 1 as training dataset.

https://doi.org/10.1371/journal.pone.0217536.g003

Discussion

This study aimed to show potential issues using a typical problematic example of retrospective multi-centric study, by investigating different methodologies and approaches for comparing cohort characteristics and assessing prognostic performance in a radiomic study. The analysis was based on a combined evaluation of four sub-cohorts, which resulted in the inclusion of 138 patients. The acquisition of a larger dataset is challenging for this research question, as FDG-PET/CT scans during treatment are not standard in clinical routine. As far as we are aware, the current dataset is one of the largest available. Nevertheless, the amount of data is probably less than required in order to build a generalized feature model for repeated PET/CT studies. Although the size of this dataset may not be sufficient from a statistical point of view, its results may provide clinical relevant insights to be obtained in order to improve future studies with limited sample sizes. Therefore, we think it is important to publish these results in an

Table 3. Model performance. Values of Harrell's concordance index with 95% confidence intervals, in the case Dataset 1 (n = 50) was used as training (T) to develop a model using LASSO. Validation (V) results are shown for Dataset 2, 3 and 4. Significant values are indicated in grey. A hyphen indicates that either all coefficients were forced to zero, or all predictions were equal to one, meaning that no linear combination of any subset of regressors was useful in predicting the outcomes.

	CT-scan1	CT-scan2	PT-scan1	PT-scan2	CT-rel	CT-abs	PT-rel	PT-abs
T-Dataset 1	**0.68**	-	-	-	**0.67**		-	**0.64**
T-Lower bound	0.59	-	-	-	0.57		-	0.56
T-Upper bound	0.76	-	-	-	0.76		-	0.73
V-Dataset 2	**0.55**	-	-	-	**0.58**		-	**0.56**
V-Lower bound	0.40	-	-	-	0.48		-	0.43
V-Upper bound	0.69	-	-	-	0.69		-	0.70
V-Dataset 3	**0.41**	-	-	-	**0.53**		-	**0.56**
V-Lower bound	0.27	-	-	-	0.39		-	0.44
V-Upper bound	0.54	-	-	-	0.68		-	0.68
V-Dataset 4	**0.49**	-	-	-	**0.62**		-	**0.52**
V-Lower bound	0.37	-	-	-	0.48		-	0.36
V-Upper bound	0.62	-	-	-	0.75		-	0.68

https://doi.org/10.1371/journal.pone.0217536.t003

attempt to guide reflections and considerations with respect to analyses and conclusions in future (radiomic) studies. In this study, both the characteristics of the study cohort and the prognostic performance of imaging features were extensively explored.

First of all, comparison of the cohort characteristics of all datasets showed that all variables were significantly different between two or more datasets, except for age. The overall stage, as well as the received radiotherapy dose of the patients in Dataset 1, was significantly different compared to the other three datasets. In addition, the timing of chemotherapy was different, since 55% of the patients from Dataset 2 received sequential chemotherapy, whereas all other patients received concurrent chemotherapy. Furthermore, the intervals between the two FDG-PET/CT scans was quite different between the datasets, since this was a retrospective

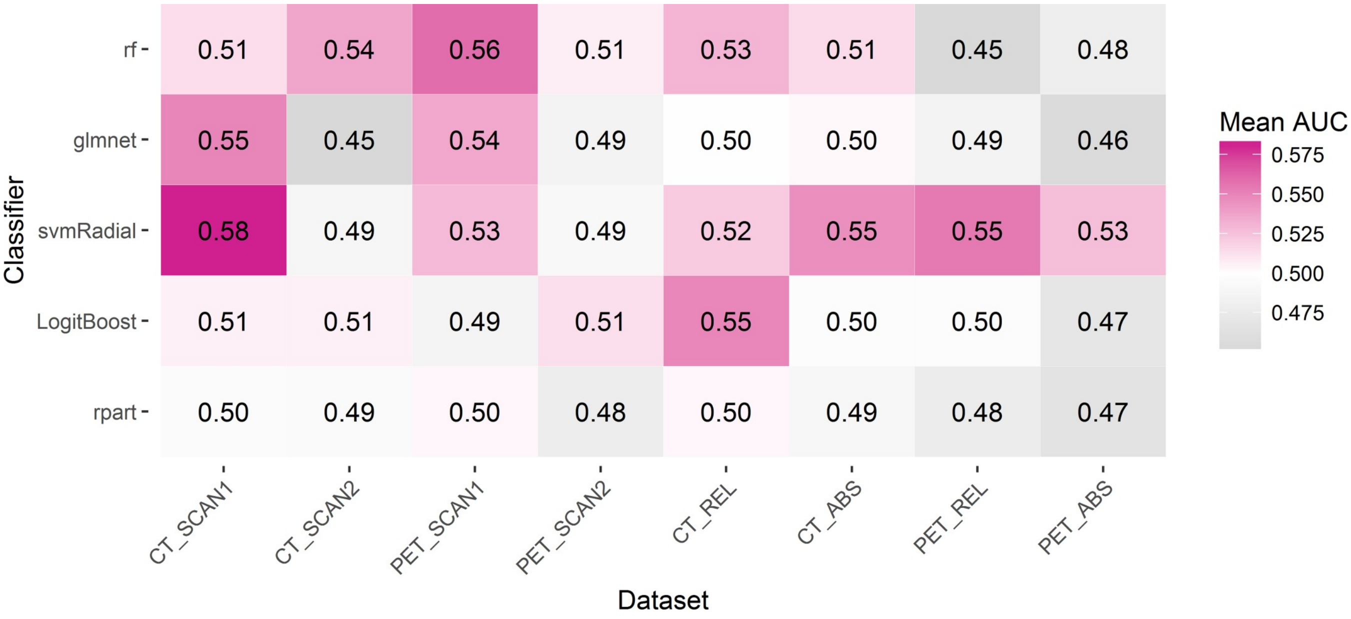

Fig 4. Classifiers heatmap. Heatmap with the mean AUC values found for each classifier and each image set, predicting two-year survival. The entire cohort with at least two-years of follow-up (n = 134) was used in this investigation.

https://doi.org/10.1371/journal.pone.0217536.g004

Table 4. Univariable analysis. Cox regression on the percentage variation of the PET imaging descriptors most commonly used, reporting the univariable hazard ratio (HR), 95% confidence interval (CI) of the HR and corresponding p-value. Univariable performance is reported in terms of the concordance-index (c-index). Absolute values of Scan 1 and 2, and percentage variation between PET acquisitions of the analyzed metrics are also presented (mean ± standard deviation).

	Scan 1	Scan 2	Percentage difference	HR	95% CI	p-value	c-index
Volume [cm^3]	85 ± 113	68 ± 94	-20 ± 31	1.11	0.55–2.25	0.77	0.50
SUV max	12 ± 6.2	9.0 ± 4.1	-20 ± 39	1.16	0.73–1.83	0.53	0.52
SUV mean	5.3 ± 2.5	4.1 ± 1.8	-12 ± 71	1.09	0.87–1.36	0.45	0.55
SUV peak	8.8 ± 4.3	6.4 ± 2.8	-18 ± 51	1.14	0.83–1.56	0.43	0.54
MTV 50%	26 ± 39	20 ± 25	-9.5 ± 46	1.72	1.09–2.71	0.02	0.54
TLG 50%	266 ± 499	132 ± 212	-25 ± 61	1.19	0.92–1.54	0.18	0.56

https://doi.org/10.1371/journal.pone.0217536.t004

study and the initial purposes of the studies were different. The variations in treatment types as well as differing outcomes between the cohorts are a probable source of noise that dilutes any signal present within the radiomic features. Furthermore, heterogeneities introduced by differences in timing of the imaging sessions, e.g. between PET scans, are a well-known confounding factor in retrospective imaging studies. Besides the large variabilities of clinical variables between the datasets, the CD-model showed high AUC values (0.65–0.88). The high AUC values of the CD-model indicate high ability to predict to which dataset the data belongs. In other words, high AUC values mean that the (distribution of) the input variables of the cohort are very different. This may be caused by differences in radiomic features since the survival is very similar for all datasets (Fig 1). Moreover, this shows that the model tests the transferability rather than the reproducibility of the model, i.e. the model is specific for a population. While the exclusion of patients with characteristics considered to be outliers would have resulted in a more homogeneous dataset, we decided against this approach as it would have resulted in too small a dataset to draw any conclusions from.

Secondly, the prognostic performance of the radiomic features extracted from CT, PET, delta-CT or delta-PET, was investigated. The models were selected using a LASSO procedure. High values of Harrell's concordance index (around 0.7 and even reaching 0.86) indicate good prognostic performance of the models trained on Dataset 2. Nevertheless, the results could not be validated on independent datasets, with the only significant c-index being 0.64 achieved on Dataset 4 for PET-scan1. This c-index was the only value being significantly different from 0.5. When one of the other datasets was used as training, no model achieved a c-index significantly different from 0.5 on any of the remaining validation datasets. In terms of discriminative power of the model, the spread of prognostic indices shown for the model trained on CT-scan1 of Dataset 1 does not show relevant outliers or differences between the datasets. In summary, the model performance results show the importance of validation: one should be careful with presenting results that are developed on (small) datasets without validation, as they are likely to be over-optimistic. In case only one (large) dataset is available, a cross-validation procedure could be applied to reduce the risk of overfitting. A recent paper summarizes the need for validation to assess the clinical usefulness of prognostic models [27].

Potentially, LASSO (e.g. penalized regression) was not the optimal method to select a prognostic model. Therefore, multiple other classifiers [28] were investigated using the methodology described by Deist *et al.* [26]. Since the sub-cohorts were small, all data were combined into one large cohort to investigate the ability of five different classifiers to predict 2-year survival. None of the classification methods was able to produce a clinical significant result, with the highest mean AUC being 0.58 for the radiomic features extracted from CT images prior to treatment. This result was achieved by the support vector machine (svm), which overall obtained the highest AUC results. Combining datasets into one large cohort did also not result in significant c-indices, potentially influenced by the heterogeneity within the combined dataset, which makes it rather difficult to develop a prognostic model.

The performance of the percentage variation of commonly used PET metrics was also investigated. Only $MTV_{50\%}$ achieved a significant Hazard Ratio, but the corresponding c-index of the Cox model was not significantly different from 0.5. This performance was also performed on the PET features extracted from either scan 1 or 2 instead of the percentage variation. None of the metrics was significantly related to overall survival (results not shown). This contradicts previously published results, which have shown the importance of these metrics in a prognostic setting [29, 30].

The results do not show conclusive results for the possibility to perform early treatment response assessment using an additional ^{18}F-FDG-PET/CT scan acquired at intermediate time point during treatment with a radiomics analysis. However, the potential prognostic value of differences between two time points was investigated before and showed promising, but also conflicting results [10, 12, 14, 16, 31]. Cremonesi *et al.* [32] reports that the differences in PET parameters are a main limitation for 21 recent studies in early response assessment for NSCLC patients. In general, radiomic features are highly affected by different acquisition and reconstruction settings [33–39]. This is one of the factors that could (partly) explain the results in the current study. In this study, there were large differences between datasets, but even within a dataset the acquisition and reconstruction settings were not identical, as shown in Table 1. The inconsistency of settings within a dataset is an issue which is difficult to overcome, especially in retrospective studies. Initiatives to provide protocols and guidelines will hopefully improve standardization in the future [40–42]. Moreover, a recent study proposed a 'post-reconstruction harmonization method' to reduce the variability in radiomic features extracted from PET images from different institutes [43]. Besides the inconsistency in acquisition and reconstruction settings within a dataset, other parts of the radiomics workflow including segmentation [44], pre-processing [45, 46] and feature extraction, remain problematic for multi-center studies and the transferability and reproducibility of developed models. A recent review proposed a harmonization for the radiomics methodology [47]. Moreover, the image biomarker standardization initiative (ISBI) attempts to standardize radiomics in terms of feature definitions and processing [48]. Nevertheless, the lack of standardization remains one the main limitations for radiomic studies. Several factors can improve the quality of future radiomic studies and reduce the risk of false positive results, as summarized in the recently proposed Radiomics Quality Score (RQS) [3].

The descriptive character of this study is intended to serve as a tool for highlighting common issues in radiomics literature and to emphasize pitfalls for future studies. Moreover, we emphasize the urgent need to publish negative results to avoid publication bias. With the current study, we would like to highlight the importance of proper validation of the results, but also the consideration of the feasibility of performing a (radiomic) study. We would recommend to perform statistically rigorous sample size calculations upfront and to perform an extensive cohort investigation to decrease the risk of false positive findings. It is essential to reduce this risk, as the main goal is to improve individual treatment decisions for better patient outcome, which can only be achieved when proper statistically sound investigations are reported.

Conclusion

In an attempt to inform future radiomic studies, we illustrate possible problems that can be encountered in retrospective multi-centric study by evaluating cohort characteristics and clinical characteristics. The models presented do not support any correlation between radiomic features acquired from FDG-PET/CT scans and overall survival. Further investigations indicate that the radiomic analysis was influenced by the limited sample size and heterogeneous imaging and clinical characteristics.

Supporting information

S1 Text. Dataset descriptions.
(DOCX)

S1 Fig. ROCs cohort difference model. Receiver Operator Curves (ROC) for the cohort difference (CD) model for each combination of datasets, including three radiomic features, 'two-year survival', 'gender' and 'stage' as independent variables.
(TIFF)

S2 Text. Feature descriptions. Mathematical feature definitions of the feature groups 'Fractal', 'Local Intensity' and 'Intensity Histogram'.
(DOCX)

S3 Text. Model performance. Harrell's concordance index with 95% confidence intervals for models trained on Dataset 2, 3, or 4 and validated on the remaining datasets.
(DOCX)

Acknowledgments

The authors would like to thank Frank Dankers and Timo Deist for their help and support using the classifier comparison methodology. The authors would also like to thank Johan van Soest for his support using the cohort differences model.

Author Contributions

Conceptualization: Janna E. van Timmeren, Sara Carvalho, Ralph T. H. Leijenaar, Esther G. C. Troost, Wouter van Elmpt, Dirk de Ruysscher, Fabrice Denis, Tanja Schimek-Jasch, Ursula Nestle, Cary Oberije.

Data curation: Janna E. van Timmeren, Sara Carvalho, Jean-Pierre Muratet, Fabrice Denis, Tanja Schimek-Jasch, Ursula Nestle.

Formal analysis: Janna E. van Timmeren.

Funding acquisition: Philippe Lambin.

Investigation: Janna E. van Timmeren, Sara Carvalho, Arthur Jochems, Henry C. Woodruff.

Methodology: Janna E. van Timmeren, Sara Carvalho, Esther G. C. Troost, Wouter van Elmpt, Arthur Jochems, Henry C. Woodruff, Cary Oberije.

Resources: Philippe Lambin.

Supervision: Wouter van Elmpt, Philippe Lambin.

Validation: Ralph T. H. Leijenaar, Esther G. C. Troost, Dirk de Ruysscher, Jean-Pierre Muratet, Cary Oberije.

Writing – original draft: Janna E. van Timmeren, Sara Carvalho.

Writing – review & editing: Janna E. van Timmeren, Sara Carvalho, Ralph T. H. Leijenaar, Esther G. C. Troost, Wouter van Elmpt, Dirk de Ruysscher, Jean-Pierre Muratet, Fabrice

Denis, Tanja Schimek-Jasch, Ursula Nestle, Arthur Jochems, Henry C. Woodruff, Cary Oberije, Philippe Lambin.

References

1. Lambin P, Zindler J, Vanneste BG, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. Advanced drug delivery reviews. 2016. Epub 2016/01/18. https://doi.org/10.1016/j.addr.2016.01.006 PMID: 26774327.
2. Lambin P, van Stiphout RG, Starmans MH, Rios-Velazquez E, Nalbantov G, Aerts HJ, et al. Predicting outcomes in radiation oncology—multifactorial decision support systems. Nat Rev Clin Oncol. 2013; 10 (1):27–40. https://doi.org/10.1038/nrclinonc.2012.196 PMID: 23165123; PubMed Central PMCID: PMCPMC4555846.
3. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017; 14(12):749–62. Epub 2017/10/05. https://doi.org/10.1038/nrclinonc.2017.141 PMID: 28975929.
4. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012; 48(4):441–6. https://doi.org/10.1016/j.ejca.2011.11.036 PMID: 22257792; PubMed Central PMCID: PMCPMC4533986.
5. Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. PloS one. 2014; 9(10):e110300. Epub 2014/10/21. https://doi.org/10.1371/journal.pone.0110300 PMID: 25330171; PubMed Central PMCID: PMCPMC4203782.
6. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012; 30(9):1234–48. https://doi.org/10.1016/j.mri.2012.06.010 PMID: 22898692; PubMed Central PMCID: PMCPMC3563280.
7. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of clinical epidemiology. 2015; 68(3):279–89. Epub 2014/09/03. https://doi.org/10.1016/j.jclinepi.2014.06.018 PMID: 25179855.
8. van Soest J, Meldolesi E, van Stiphout R, Gatta R, Damiani A, Valentini V, et al. Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility. Medical physics. 2017; 44(9):4961–7. Epub 2017/06/24. https://doi.org/10.1002/mp.12423 PMID: 28639302.
9. 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(5):e0124165. Epub 2015/05/06. https://doi.org/10.1371/journal.pone.0124165 PMID: 25938522; PubMed Central PMCID: PMCPMC4418696.
10. Usmanij EA, de Geus-Oei LF, Troost EG, Peters-Bax L, van der Heijden EH, Kaanders JH, et al. 18F-FDG PET early response evaluation of locally advanced non-small cell lung cancer treated with concomitant chemoradiotherapy. Journal of nuclear medicine: official publication, Society of Nuclear Medicine. 2013; 54(9):1528–34. Epub 2013/07/19. https://doi.org/10.2967/jnumed.112.116921 PMID: 23864719.
11. Massaccesi M, Calcagni ML, Spitilli MG, Cocciolillo F, Pelligro F, Bonomo L, et al. (1)(8)F-FDG PET-CT during chemo-radiotherapy in patients with non-small cell lung cancer: the early metabolic response correlates with the delivered radiation dose. Radiation oncology (London, England). 2012; 7:106. Epub 2012/07/12. https://doi.org/10.1186/1748-717x-7-106 PMID: 22781363; PubMed Central PMCID: PMCPMC3410758.
12. van Elmpt W, Ollers M, Dingemans AM, Lambin P, De Ruysscher D. Response assessment using 18F-FDG PET early in the course of radiotherapy correlates with survival in advanced-stage non-small cell lung cancer. Journal of nuclear medicine: official publication, Society of Nuclear Medicine. 2012; 53 (10):1514–20. Epub 2012/08/11. https://doi.org/10.2967/jnumed.111.102566 PMID: 22879081; PubMed Central PMCID: PMCPMC4667805.
13. Huang W, Zhou T, Ma L, Sun H, Gong H, Wang J, et al. Standard uptake value and metabolic tumor volume of (1)(8)F-FDG PET/CT predict short-term outcome early in the course of chemoradiotherapy in advanced non-small cell lung cancer. European journal of nuclear medicine and molecular imaging. 2011; 38(9):1628–35. Epub 2011/05/28. https://doi.org/10.1007/s00259-011-1838-5 PMID: 21617977.
14. Banna GL, Anile G, Russo G, Vigneri P, Castaing M, Nicolosi M, et al. Predictive and Prognostic Value of Early Disease Progression by PET Evaluation in Advanced Non-Small Cell Lung Cancer. Oncology. 2017; 92(1):39–47. Epub 2016/11/11. https://doi.org/10.1159/000448005 PMID: 27832654.
15. Hofheinz F, Hoff J, Steffen IG, Lougovski A, Ego K, Amthauer H, et al. Comparative evaluation of SUV, tumor-to-blood standard uptake ratio (SUR), and dual time point measurements for assessment of the

metabolic uptake rate in FDG PET. EJNMMI research. 2016; 6(1):53. Epub 2016/06/24. https://doi.org/10.1186/s13550-016-0208-5 PMID: 27334609; PubMed Central PMCID: PMCPMC4917506.

16. Yossi S, Krhili S, Muratet JP, Septans AL, Campion L, Denis F. Early assessment of metabolic response by 18F-FDG PET during concomitant radiochemotherapy of non-small cell lung carcinoma is associated with survival: a retrospective single-center study. Clinical nuclear medicine. 2015; 40(4): e215–21. Epub 2014/12/30. https://doi.org/10.1097/RLU.0000000000000615 PMID: 25546211.

17. Kinahan PE, Fletcher JW. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. Seminars in ultrasound, CT, and MR. 2010; 31(6):496–505. Epub 2010/12/15. https://doi.org/10.1053/j.sult.2010.10.001 PMID: 21147377; PubMed Central PMCID: PMCPMC3026294.

18. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta oncologica (Stockholm, Sweden). 2017; 56(11):1544–53. Epub 2017/09/09. https://doi.org/10.1080/0284186x.2017.1351624 PMID: 28885084.

19. van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiotherapy and oncology: journal of the European Society for Therapeutic Radiology and Oncology. 2017; 123(3):363–9. Epub 2017/05/17. https://doi.org/10.1016/j.radonc.2017.04.016 PMID: 28506693.

20. Van de Wiele C, Kruse V, Smeets P, Sathekge M, Maes A. Predictive and prognostic value of metabolic tumour volume and total lesion glycolysis in solid tumours. European journal of nuclear medicine and molecular imaging. 2013; 40(2):290–301. Epub 2012/11/16. https://doi.org/10.1007/s00259-012-2280-z PMID: 23151913.

21. Visser EP, Philippens ME, Kienhorst L, Kaanders JH, Corstens FH, de Geus-Oei LF, et al. Comparison of tumor volumes derived from glucose metabolic rate maps and SUV maps in dynamic 18F-FDG PET. Journal of nuclear medicine: official publication, Society of Nuclear Medicine. 2008; 49(6):892–8. Epub 2008/05/17. https://doi.org/10.2967/jnumed.107.049585 PMID: 18483085.

22. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2 ed: Springer-Verlag New York; 2008. 745 p.

23. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267–88.

24. Harrell FE Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982; 247(18):2543–6. Epub 1982/05/14. PMID: 7069920.

25. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013; 13:33. Epub 2013/03/19. https://doi.org/10.1186/1471-2288-13-33 PMID: 23496923; PubMed Central PMCID: PMCPMC3667097.

26. Deist TM, Dankers F, Valdes G, Wijsman R, Hsu IC, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Medical physics. 2018. Epub 2018/05/16. https://doi.org/10.1002/mp.12967 PMID: 29763967.

27. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiotherapy and Oncology. https://doi.org/10.1016/j.radonc.2018.03.004 PMID: 29598835

28. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015; 13:8–17. Epub 2015/03/10. https://doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696; PubMed Central PMCID: PMCPMC4348437.

29. Paesmans M, Garcia C, Wong CY, Patz EF Jr., Komaki R, Eschmann S, et al. Primary tumour standardised uptake value is prognostic in nonsmall cell lung cancer: a multivariate pooled analysis of individual data. Eur Respir J. 2015; 46(6):1751–61. Epub 2015/09/26. https://doi.org/10.1183/13993003.00099-2015 PMID: 26405289.

30. Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, et al. Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors. Radiology. 2016; 278(1):214–22. Epub 2015/07/16. https://doi.org/10.1148/radiol.2015142920 PMID: 26176655; PubMed Central PMCID: PMCPMC4699494.

31. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific reports. 2017; 7(1):588. Epub 2017/04/05. https://doi.org/10.1038/s41598-017-00665-z PMID: 28373718.

32. Cremonesi M, Gilardi L, Ferrari ME, Piperno G, Travaini LL, Timmerman R, et al. Role of interim (18)F-FDG-PET/CT for the early prediction of clinical outcomes of Non-Small Cell Lung Cancer (NSCLC) during radiotherapy or chemo-radiotherapy. A systematic review. European journal of nuclear medicine and molecular imaging. 2017; 44(11):1915–27. Epub 2017/07/07. https://doi.org/10.1007/s00259-017-3762-9 PMID: 28681192.

33. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta oncologica (Stockholm, Sweden). 2010; 49(7):1012–6. Epub 2010/09/14. https://doi.org/10.3109/0284186x.2010.498437 PMID: 20831489; PubMed Central PMCID: PMCPmc4091820.

34. van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. Mol Imaging Biol. 2016. https://doi.org/10.1007/s11307-016-0940-2 PMID: 26920355.

35. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. Journal of nuclear medicine: official publication, Society of Nuclear Medicine. 2015; 56(11):1667–73. https://doi.org/10.2967/jnumed.115.156927 PMID: 26229145.

36. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. Journal of medical imaging (Bellingham, Wash). 2015; 2(4). https://doi.org/10.1117/1.jmi.2.4.041002 PMID: 26251842; PubMed Central PMCID: PMCPMC4524811.

37. Mackin D, Ger R, Dodge C, Fave X, Chi P-C, Zhang L, et al. Effect of tube current on computed tomography radiomic features. Scientific reports. 2018; 8(1):2354. https://doi.org/10.1038/s41598-018-20713-6 PMID: 29403060

38. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring Computed Tomography Scanner Variability of Radiomics Features. Investigative radiology. 2015; 50(11):757–65. https://doi.org/10.1097/RLI.0000000000000180 PMID: 26115366; PubMed Central PMCID: PMCPMC4598251.

39. Bailly C, Bodet-Milin C, Couespel S, Necib H, Kraeber-Bodere F, Ansquer C, et al. Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. PloS one. 2016; 11 (7):e0159984. Epub 2016/07/29. https://doi.org/10.1371/journal.pone.0159984 PMID: 27467882; PubMed Central PMCID: PMCPMC4965162.

40. Boellaard R, O'Doherty MJ, Weber WA, Mottaghy FM, Lonsdale MN, Stroobants SG, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. European journal of nuclear medicine and molecular imaging. 2010; 37(1):181–200. Epub 2009/11/17. https://doi.org/10.1007/s00259-009-1297-4 PMID: 19915839; PubMed Central PMCID: PMCPMC2791475.

41. Boellaard R, Delgado-Bolton R, Oyen WJ, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. European journal of nuclear medicine and molecular imaging. 2015; 42(2):328–54. Epub 2014/12/03. https://doi.org/10.1007/s00259-014-2961-x PMID: 25452219; PubMed Central PMCID: PMCPMC4315529.

42. de Jong EEC, van Elmpt W, Hoekstra OS, Groen HJM, Smit EF, Boellaard R, et al. Quality assessment of positron emission tomography scans: recommendations for future multicentre trials. Acta oncologica (Stockholm, Sweden). 2017; 56(11):1459–64. Epub 2017/08/24. https://doi.org/10.1080/0284186x.2017.1346824 PMID: 28830270.

43. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. Journal of nuclear medicine: official publication, Society of Nuclear Medicine. 2018. Epub 2018/01/06. https://doi.org/10.2967/jnumed.117.199935 PMID: 29301932.

44. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PloS one. 2014; 9(7):e102107. https://doi.org/10.1371/journal.pone.0102107 PMID: 25025374; PubMed Central PMCID: PMCPMC4098900.

45. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Scientific reports. 2015; 5:11075. https://doi.org/10.1038/srep11075 PMID: 26242464; PubMed Central PMCID: PMCPMC4525145.

46. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PloS one. 2017; 12(9):e0178524. Epub 2017/09/22. https://doi.org/10.1371/journal.pone.0178524 PMID: 28934225; PubMed Central PMCID: PMCPMC5608195.

47. Sollini M, Cozzi L, Antunovic L, Chiti A, Kirienko M. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. Scientific reports. 2017; 7(1):358. Epub 2017/03/25. https://doi.org/10.1038/s41598-017-00426-y PMID: 28336974; PubMed Central PMCID: PMCPMC5428425.

48. Zwanenburg A, Leger S, Vallieres M, Lock S. Image biomarker standardisation intitiative—feature definitions. CoRR. 2016;abs/1612.07003.