European Journal of Radiology 113 (2019) 101–109

Contents lists available at ScienceDirect

European Journal of Radiology

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

Research article

Prognostic modeling for patients with colorectal liver metastases incorporating FDG PET radiomic features

Arman Rahmim[a,b,*], Kirstine P. Bak-Fredslund[c], Saeed Ashrafinia[a,d], Lijun Lu[e], C. Ross Schmidtlein[f], Rathan M. Subramaniam[g], Anni Morsing[c], Susanne Keiding[c], Jacob Horsager[c], Ole L. Munk[c]

[a] Department of Radiology and Radiological Science, Johns Hopkins University, Baltimore, MD, USA
[b] Departments of Radiology and Physics & Astronomy, University of British Columbia, Vancouver, BC, Canada
[c] Department of Nuclear Medicine and PET Center, Aarhus University Hospital, Aarhus, Denmark
[d] Department of Electrical & Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
[e] School of Biomedical Engineering, Southern Medical University, Guangzhou, China
[f] Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[g] Department of Radiology, University of Texas Southwestern Medical Center, TX, USA

ARTICLE INFO

Keywords:
PET/CT
Colorectal liver metastasis
Prognosis
Volumetric features
Intra-tumoral heterogeneity
Radiomics

ABSTRACT

Objective: We aimed to improve prediction of outcome for patients with colorectal liver metastases, via prognostic models incorporating PET-derived measures, including radiomic features that move beyond conventional standard uptake value (SUV) measures.

Patients and methods: A range of parameters including volumetric and heterogeneity measures were derived from FDG PET images of 52 patients with colorectal intrahepatic-only metastases (29 males and 23 females; mean age 62.9 years [SD 9.8; range 32–82]). The patients underwent PET/CT imaging as part of the clinical workup prior to final decision on treatment. Univariate and multivariate models were implemented, which included statistical considerations (to discourage false discovery and overfitting), to predict overall survival (OS), progression-free survival (PFS) and event-free survival (EFS). Kaplan-Meier survival analyses were performed, where the subjects were divided into high-risk and low-risk groups, from which the hazard ratios (HR) were computed via Cox proportional hazards regression.

Results: Commonly-invoked SUV metrics performed relatively poorly for different prediction tasks (SUVmax HR = 1.48, 0.83 and 1.16; SUVpeak HR = 2.05, 1.93, and 1.64, for OS, PFS and EFS, respectively). By contrast, the number of liver metastases and metabolic tumor volume (MTV) each performed well (with respective HR values of 2.71, 2.61 and 2.42, and 2.62, 1.96 and 2.29, for OS, PFS and EFS). Total lesion glycolysis (TLG) also resulted in similar performance as MTV. Multivariate prognostic modeling incorporating different features (including those quantifying intra-tumor heterogeneity) resulted in further enhanced prediction. Specifically, HR values of 4.29, 4.02 and 3.20 (p-values = 0.00004, 0.0019 and 0.0002) were obtained for OS, PFS and EFS, respectively.

Conclusions: PET-derived measures beyond commonly invoked SUV parameters hold significant potential towards improved prediction of clinical outcome in patients with liver metastases, especially when utilizing multivariate models.

1. Introduction

Colorectal cancer is a common cancer worldwide, often burdened by liver metastases [1]. About 15% of patients have liver metastases at the time of diagnosis and an additional 15% developed liver metastases over time [2]; 5-year survival in patients with liver metastases was reported as low as 5% in untreated patients [2]. However, recent studies report a 5-year survival rate of about 40% following surgical resection of colorectal liver metastases [3]. Treatment options for colorectal liver metastases have expanded with new therapeutic modalities such as radiofrequency ablation, which imply a clinical need for improved prognostication to assist choice of therapy.

[*] Corresponding author at: University of British Columbia, Vancouver, BC Cancer Research Centre, 675 West 10th Ave.; Office 5-114, BC, V5Z 1L3, Canada.
E-mail address: arman.rahmim@ubc.ca (A. Rahmim).

https://doi.org/10.1016/j.ejrad.2019.02.006
Received 23 March 2018; Received in revised form 22 December 2018; Accepted 4 February 2019

The emerging area of precision (or personalized) cancer medicine involves efforts towards the discovery and validation of biomarkers that move beyond diagnosis, to domains such as prognostication, disease progression tracking, and therapy response prediction and assessment. To this end, PET imaging provides valuable capabilities for non-invasive assessment and quantification of disease burden, and towards the development of effective imaging biomarkers of disease [4]. Overall, PET images present a wide array of information related to disease. However, in common clinical practice, only intensity-based standard-uptake-value (SUV) metrics are utilized, particularly SUVmax or SUVpeak. This is due to the simplicity in the computation of these metrics, not requiring accurate segmentation of the tumors. Specifically, SUVmax is computed as the maximum uptake in an area of interest, and SUVpeak is obtained by moving a 1-cm^3 spherical region of interest over the area with increased tracer uptake (not necessarily conforming to the precise tumor outline) to maximize the enclosed average uptake [5,6].

Quantitative volumetric tumor parameters, though less straightforward to compute, provide a notable frontier towards improved assessment of disease. In fact, there is increasing evidence that volumetric measures, particularly metabolic tumor volume (MTV) or total lesion glycolysis (TLG) can outperform their SUV counterparts, in a range of human solid tumors such as head & neck cancer, lung cancer, breast cancer, colorectal cancer and lymphoma [7–16]. Tumor volumetric parameters facilitate estimation of total tumor burden in a patient at the time of diagnosis or recurrence. Furthermore, segmentation of PET images enables generation of SUVmean, which is also sometimes reported in the literature.

In the present work, we have performed extensive comparisons, including univariate and multivariate analyses involving a range of quantitative measures of tumor uptake, to assess optimal methods for prediction of clinical outcome in patients with liver metastases from colorectal cancer. Our analyses includes the use of volumetric parameters, as well as other advanced radiomic features which quantify heterogeneity [17–21] as increasingly studied in the emerging field of radiomics. The ultimate aim is that enhanced predictive models would result in significant improvements in management of patients, including non-invasive selection of patients with poor prognosis who could benefit from earlier and more intensive treatment strategies. These high-risk patients could also be identified for participation in clinical trials in order to better power discovery of effective therapies.

2. Patients and methods

2.1. Subjects

We analyzed data from 52 patients with colorectal intrahepatic-only metastases (29 males and 23 females; mean age 62.9 years [SD 9.8; range 32–82]). The patients had FDG PET/CT scans obtained before treatment, in years 2005 to 2010 (with patient outcome follow-ups up to 2017). The scans were performed as part of the clinical workup prior to final decision on treatment, most often in patients considered for liver surgery, as PET/CT was not part of primary standard workup for all patients with liver metastases from colorectal cancer. Treatment for liver metastases following FDG PET/CT included surgical resection, stereotactic radiotherapy, chemotherapy, radiofrequency ablation, or a combination of these therapies. The treatment modalities were modeled in our analyses.

We performed analyses of overall survival (OS), progression-free survival (PFS) and event-free survival (EFS) for imaging biomarker derivation. Progression was defined as local recurrence in the liver, or new metastases in the liver or outside the liver. This could include new tumors in the intestine detected with ordinary control examinations: mainly, contrast-enhanced CT of the thorax, abdomen and pelvis, and in few cases MRI and ultrasound. Of the 52 patients, number of events for OS (death), PFS (progression) and EFS (progression or death) were 40, 25 and 44, respectively. The PET/CT scans were acquired on Siemens Biograph TruePoint scanners at the PET Centre of Aarhus University Hospital. Typical acquisitions started at 60 min post-injection, from top of head to mid-thigh, and spanned 3 min/bed. Reconstructions involved iterative 2D OSEM which was chosen for consistency amongst patients including those scanned in earlier years (see discussion section).

2.2. Data analysis

Segmentation: Tumors were segmented based on the PET images, though the fused PET/CT images were used initially to ensure that the tumors were intrahepatic (and not metastases in lung or peritoneum). The identified tumors were segmented using: (i) 40% background-corrected SUVmax, (ii) 50% background-corrected SUVmax, (iii) SUV > 2.5, or (iv) SUV > 3.0 thresholding, all in 3D using the Hermes Hybrid Viewer PDR software (Hermes Medical Solutions, Sweden). Background correction was performed using a liver background ROI (˜14 mL) placed on liver tissue with good distance to tumors, followed by contouring based on t = 40% or 50% lower threshold, calculated as [SUVmax(tumor) – SUVmean(background)] × t + SUVmean(background) [22]. Histograms of PET counts were generated from the segmented tumors (in increments of ˜0.02 SUV units used for creating discretized gray levels). This allowed moving beyond conventional PET-derived measures and to generate radiomic features quantifying heterogeneity (as elaborated next). In patients with multiple liver metastases (average of 1.8 tumors/person; 21 patients with multiple metastases), the histograms were combined, and subsequently analyzed.

Data features: A total of 51 features were extracted from each patient. This included 41 image-derived radiomic features (as described in the next paragraph), and 10 features as follows: (1) age, (2) sex, and (3) post-imaging treatment information (described earlier, and modeled as input features in our analyses). We also incorporated pre-imaging treatment information, such as whether any therapy was delivered to the liver: (4) prior to PET (liver-therapy-prior), or (5) < 3 months prior to PET (liver-therapy-3mon-prior), or whether chemotherapy itself was specifically performed (6) prior to PET (chemotherapy-prior), or (7) < 3 months prior to PET (chemotherapy-3mon-prior). We also included (8) number of liver metastases observed in PET scan. Furthermore, we categorized patients based on (9) whether metastases were detected by the time of diagnosis (synchronous) vs. up to 12 months after diagnosis (early metachronous) vs. more than 12 months after diagnosis (late metachronous) [3]. We also explored another categorization for condition of existing metastases: (10) whether metastases were absent by the time of diagnosis (metachronous) vs. present at diagnosis, this latter itself consisting of two subsets: whether the *specific* tumors visualized by existing PET scan were present vs. absent at time of primary diagnosis.

We extracted 41 quantitative imaging features (radiomic features), which are elaborated in supplement A. To summarize, we included SUVmax, SUVpeak, SUVmean, MTV and TLG (thus n = 5). We also computed a range of radiomic features that quantified PET-uptake heterogeneity. This included the recently introduced class of generalized effective total uptake (gETU) measures [23] which place varying degrees of emphasis on volumetric vs. uptake information (n = 10), which are further discussed in the discussion Section 4.2. It also included intensity histogram (n = 19) and intensity-volume histogram (IVH) (n = 7) measures [24,25]. All metrics used in this work were standardized according to the framework of the image biomarker standardization initiative (IBSI) [26] for wider applicability of our results to other users and centers. In the results section, we report on the performance of SUVmax, SUVpeak, SUVmean, MTV and TLG, as well as any other metrics that were found to be significant in univariate or multivariate analyses.

Survival analysis: Kaplan-Meier survival analysis was performed for

OS, PFS and EFS, including both univariate and multivariate analyses. Prior to performing these analyses, feature selection was performed. Spearman correlations (*r*) amongst the 51 measures were computed, and those with $r > 0.95$ were considered relatively redundant with respect to one another (the results were nearly identical with the use of Pearson correlations). Subsequently, we reduced the original list to a narrow list. This was followed by application of (a) univariate and (b) multivariate survival analyses, which included statistical methods specific to each, as elaborated next, and as implemented in-house using MATLAB software.

a) Univariate analysis: The subjects were subdivided into two groups using the median threshold (p = 50th percentile) for a given metric (e.g. MTV, etc.). Following this, the hazard ratios (HR) between the higher percentile group to the lower percentile groups were computed using Cox proportional hazards regression, and their associated 95% confidence intervals (CI) were also derived. For each metric, we also computed the p-values for curve separation (i.e. ability to reject the null hypothesis that HR = 1). Correction for multiple testing of different metrics was performed using the false discovery rate (FDR) Benjamini–Hochberg (BH) step-up procedure.
b) Multivariate analysis: Cox proportional hazards regression was again performed. A prognostic score was then generated for each multivariate Cox model by summing the products of each feature in the model and its corresponding regression coefficient (β). The median value of the prognostic score was then chosen as cut-off for the given model, and patients were thus dichotomized into low- and high-risk groups, for which the log-likelihood (LOGL) of Cox regression was measured. Stepwise forward selection of parameters was performed. Specifically, we tried two initializations: a model with a single metric that outperformed others in univariate analysis, or with a single conventional metric that outperformed others (see discussion). Subsequently we would test the inclusion of every metric that was not in the model, adding to the model the one that most significantly increased LOGL as quantified above. This process was repeated as long as addition of a new metric increased LOGL statistically significantly. Statistical significance between two models was assessed using the Akaike Information Criterion (AIC) for model selection, which would require an increase in LOGL by > 1 by addition of a new metric. This constraint was imposed on model selection in order to discourage overfitting. In the discussion section, we discuss the use of more stringent criteria.

3. Results

An example of segmentation for a subject with liver metastasis is depicted in Fig. 1. When using SUV metrics, the four segmentation methods performed relatively similarly, but when performing volumetric analysis, 40% and 50% background-corrected SUVmax thresholding resulted in relatively improved performance especially in PFS (elaborated in the discussion section). Rest of the paper describes results for 40% background-corrected SUVmax thresholding.

Of the original 51 metrics, 26 were retained following correlation analysis (listed in supplement B). We note that SUVmax and SUVmean were found to be highly correlated with SUVpeak (r = 0.98, p-value < 0.0001 for both). However, they were retained for further analysis and reporting (to allow comparison with prior literature).

Subsequently, univariate and multivariate Cox regression analyses were performed, and the respective results are summarized in Tables 1 and 2, which indicate HR values, their associated 95% CI and p-values (i.e. of rejecting the null hypothesis that HR = 1). In addition, the performances are visually depicted using Kaplan-Meier plots in Figs. 2–4 for OS, PFS and EFS, respectively.

The univariate results for number of liver mets, MTV, TLG, SUVpeak, SUVmean and SUVmax are specifically summarized in Table 1 (presented in order of significance), and plotted in Figs. 2–4. We found that the number of liver mets, MTV and TLG outperformed the other metrics for OS, PFS and EFS. Amongst these variables, survival discrimination (p-value) was only significant for MTV and number of liver mets in both cases of OS and EFS after correction for multiple testing, as indicated in Table 1. It is also notable to see that amongst the metrics listed in Table 1, SUVmax (most commonly reported PET metric) performs the most poorly, and that volumetric measures perform better.

The radiomic feature V_{10-90} had p-values < 0.05 in PFS and EFS analyses (0.014 and 0.025, respectively). It is elaborated in supplement A; in short, it is an IVH based metric, quantifying the difference between fraction of volume of the segmented tumor with intensities at least 10% (V_{10}) and 90% (V_{90}) of maximum gray level (i.e. $V_{10\text{-}90} = V_{10}\text{-}V_{90}$). Also, whether any therapy was delivered to the liver < 3 months prior to PET (liver-therapy-3mon-prior) had p-values of 0.021 for OS. However, performance of these features was not significant after correction for multiple testing.

Subsequently, we performed multivariate analysis, using stepwise Cox regression (forward selection) as elaborated in the methods section. The results are summarized in Table 2. In the case of OS, HR value of 4.29 was obtained (also depicted in Fig. 2). The final multivariate model (arrived at according to the statistical methods described in the

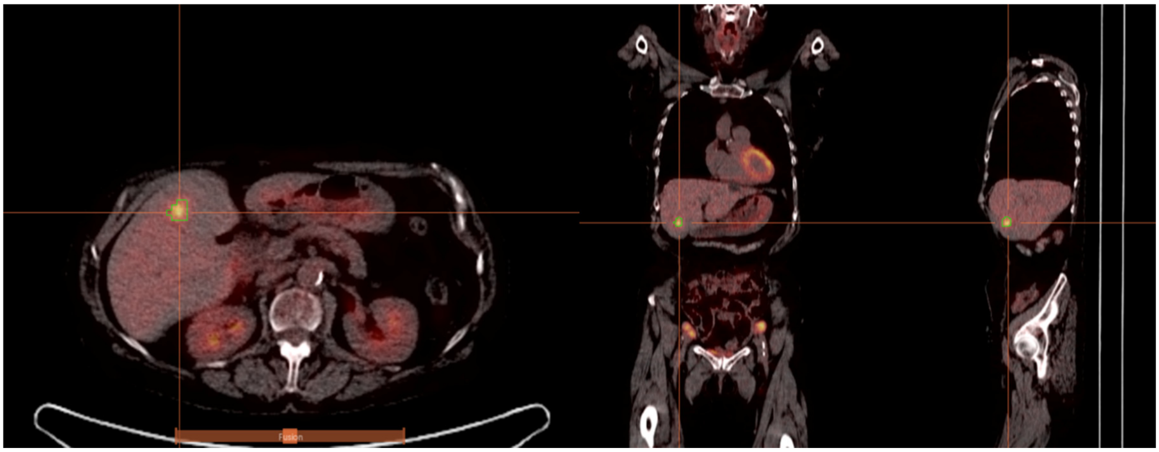

Fig. 1. PET/CT images from a patient with colorectal liver metastasis. Sagittal, coronal, and transaxial slices (left-to-right) are shown. Overlaid on the images is 3D PET-derived segmentation via 40% background-corrected SUVmax thresholding. Tumors were identified based on the PET images, though the fused PET/CT images were used initially to ensure that the tumors were intrahepatic (and not metastases in lung or peritoneum).

Table 1
Univariate Cox Regression Analysis for OS, PFS and EFS.

	OS		PFS		EFS	
Parameters*	HR (95% CI)	p-value	HR (95% CI)	p-value	HR (95% CI)	p-value
Num. of liver mets	2.71 (1.44-5.12)	0.0021**	2.61 (1.18-5.79)	0.018	2.42 (1.32-4.42)	0.0042
MTV	2.62 (1.38-4.98)	0.0034**	1.96 (0.87-4.41)	0.11	2.29 (1.23-4.24)	0.0086
TLG	2.62 (1.38-4.98)	0.0034**	1.96 (0.87-4.41)	0.11	2.29 (1.23-4.24)	0.0086
SUVpeak	2.05 (1.09-3.86)	0.027	1.93 (0.86-4.33)	0.11	1.64 (0.90-2.99)	0.10
SUVmean	1.81 (0.96-3.41)	0.068	0.82 (0.37-1.80)	0.62	1.35 (0.74-2.44)	0.33
SUVmax	1.48 (0.79-2.77)	0.22	0.83 (0.38-1.82)	0.64	1.16 (0.64-2.09)	0.63

* Median thresholds for MTV, TLG, SUVpeak, SUVmean and SUVmax were 9.3 mL, 58.3 mL, 6.8, 5.3 and 7.8, respectively, arriving at 26 patients in each of the lower and higher risk groups. Number of liver mets was set to = 1 vs. > 1 arriving at 31 vs. 21 patients in the lower and higher risk groups.

** p-values significant after correction for multiple testing (to control for FDR).

Table 2
Multivariate Cox Regression Analysis for OS, PFS and EFS.

Survival Analysis	Parameters in the model	HR (95% CI)	p-value
OS	Num. of liver mets Liver-therapy-3mon-prior AUC-IVH	4.29 (2.15-8.57)	0.00004
PFS	Num. of liver mets SUVmax	4.02 (1.67-9.70)	0.0019
EFS	Num. of liver mets MTV Histogram uniformity	3.20 (1.73-5.94)	0.0002

methods section) included three metrics, namely (i) number of liver mets, (ii) liver-therapy-3mon-prior, and (iii) AUC-IVH. The definitions of radiomic features are provided in supplement A; in short, AUC-IVH is the area under the IVH curve, also known as AUC-CSH [25], which quantifies tumoral heterogeneity. In the case of PFS, an HR value of 4.02 was obtained (also depicted in Fig. 3) where the final model included (i) number of liver mets, and (ii) SUVmax. Finally, in the case of EFS, an HR value of 3.20 was obtained (also depicted in Fig. 4), and the final multivariate model consisted of three metrics, namely (i) number of liver mets, (ii) MTV, and (iii) histogram uniformity (also known as histogram energy) which is computed as the sum of squares of occurrence probabilities of discretized histogram intensities (see supplement A). Note that it is possible for parameters not to be significant in univariate analysis but to become significant in multivariate analysis [27].

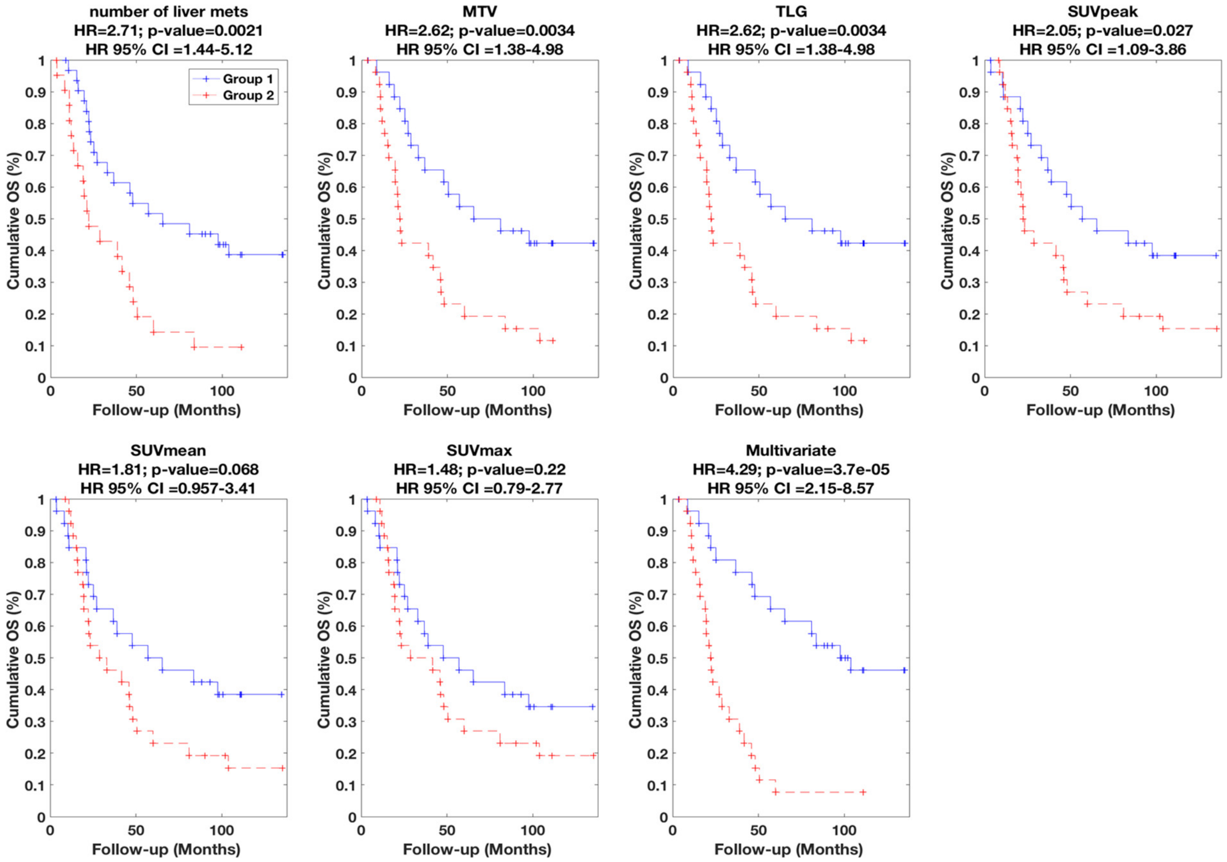

Fig. 2. Kaplan-Meier curves for OS. Univariate results are shown for six different metrics (number of liver mets, MTV, TLG, SUVpeak, SUVmean, SUVmax), while multivariate result is also shown (Table 2 lists parameters in model). (+) signs indicate events at steps or last follow-up otherwise. Segmentation was performed using 40% background-corrected SUVmax thresholding. HR as well as associated 95%-CI and p-values are also reported. Group 1 (lower risk) vs. Group 2 (higher risk) had 26 vs. 26 subjects in all plots, except for univariate *number of liver mets* (31 vs. 21).

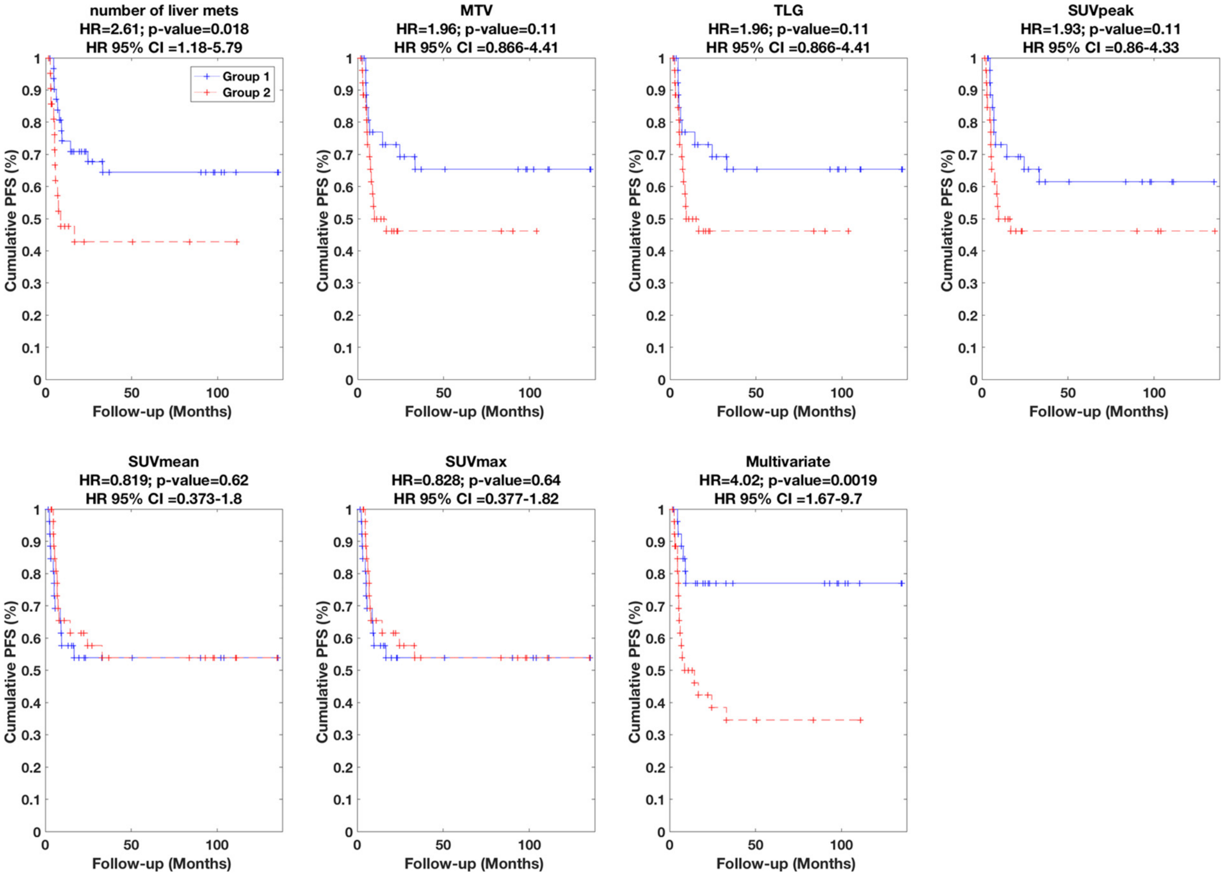

Fig. 3. Kaplan-Meier curves for PFS. Univariate results are shown for six different metrics, while multivariate result is also shown (Table 2 lists parameters in model).

When building prognostic models excluding the use of volumetric or heterogeneity features (i.e. only using features 1–12 in supplement B), HR for OS dropped from 4.29 to 3.77 (with features: number of liver mets and liver-therapy-3mon-prior), while decreasing from 3.20 to 2.42 for EFS (with feature: number of liver mets). HR for PFS remained the same. When further excluding any imaging features (i.e. making no use of images), models with only a single metric were obtained, with HR values of 2.34 for OS (chemotherapy-3mon-prior), 2.08 for PFS (sex) and 1.82 for EFS (chemotherapy-3mon-prior).

Overall, it was seen that a simple imaging feature, namely the number of liver mets, performed strongly in univariate prediction of outcome, in contrast to SUV measures especially SUVmax, which did not perform well. Volumetric measures of MTV and TLG also depicted significant performance. Moreover, multivariate prognostic models incorporating radiomic features further improved prediction of outcome. Consequently, it was seen that volumetric and/or heterogeneity features that move beyond conventional SUV measures have the potential for significant prediction of outcome in patients with colorectal liver metastases.

4. Discussion

4.1. *Conventional measures vs. volumetric and heterogeneity parameters*

In our univariate survival analyses of OS, PFS and EFS, SUV measures (max/mean/peak) did not perform as well as volumetric measures MTV or TLG (Figs. 2–4). Furthermore, in multivariate analyses, only in the case of PFS, SUV added value in combination with number of liver mets.

In a study by de Geus-Oei et al. [28] of 152 colorectal metastatic patients (majority with involvement of the liver), only SUVmean was evaluated for OS. The resulting HR, though statistically significant, was only 1.17, while it was 1.81 in our study (Table 1). By contrast, SUVmax was evaluated for both OS and PFS by Dimitrova et al. [29] in a study of 43 patients with colon cancer and unresectable liver metastases. SUVmax was not able to predict PFS, though it predicted OS with HR value of 2.05 (while it was 1.48 in our study). In a study by De Bryne et al. [30] of 19 metastatic colorectal cancer patients with potentially resectable liver lesions, post-treatment SUVmax was only reported, and an HR value of 1.20 was obtained for prediction of PFS that was not statistically significant. The key finding in our analyses is that volumetric and heterogeneity metrics beyond SUV hold value for improved predictions of outcome.

Vriens et al. [31] evaluated 23 patients with colorectal liver metastases. The subjects underwent dynamic PET imaging, followed by measurement of glucose metabolic rates (MRglc) via Patlak graphical analysis. The authors demonstrated significant performances for OS (HR = 3.61) and PFS (HR = 3.11). It is unclear, and remains to be seen, how volumetric or heterogeneity features would perform in comparison if applied in the domain of dynamic PET imaging. The dynamic scans spanned a total of 50 min from time of injection, and thus performance for routine static imaging (typically at 60 min post-injection) was not reported in the study. Usage of dynamic scanning is expected to remain limited in the wide clinical setting which commonly employs whole-body imaging. An alternative paradigm worth exploring is to incorporate dynamic imaging within multi-bed/whole-body imaging [32,33].

Gulec et al. [34] and Shady et al. [35] studied 20 and 49 patients, respectively, undergoing ^{90}Y radioembolization of colorectal liver metastasis. Both studies reported ability of MTV and TLG measures to predict OS. In the former study, this was shown for pre- and post-treatment scans individually, while comparison with conventional SUV measures was not reported. In the latter study, by contrast, response measures (i.e. changes from pre- to post-treatment scans) were used,

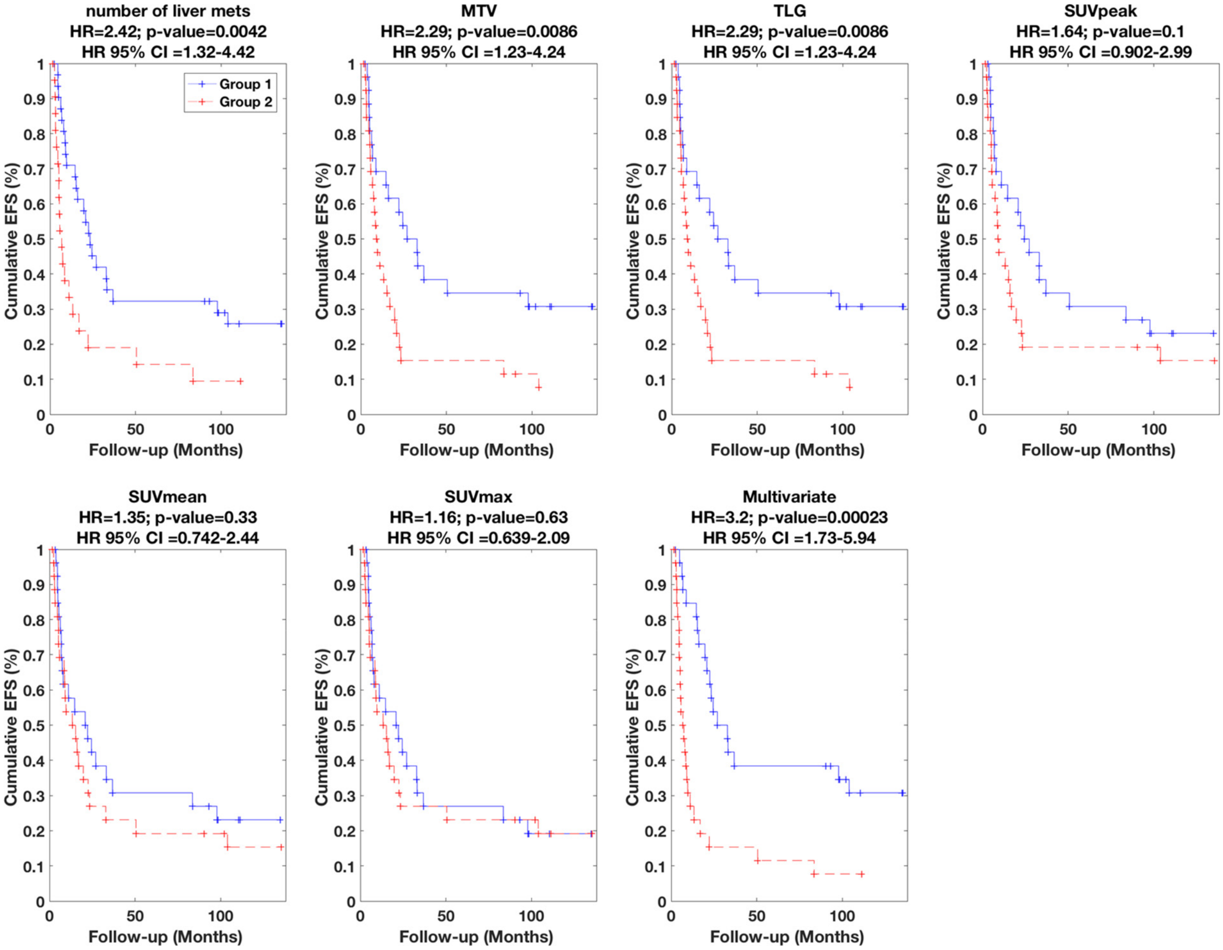

Fig. 4. Kaplan-Meier curves for EFS. Univariate results are shown for six different metrics, while multivariate result is also shown (Table 2 lists parameters in model).

and it was additionally shown that response measures by SUVmax and SUVpeak were not predictive of OS, nor was CT-based Response Evaluation Criteria In Solid Tumors (RECIST) 1.0. By contrast, in a study by Lastoria et al. [36] of 33 colorectal cancer patients with resectable liver metastases, response assessments by both SUVmax and TLG were found to add value to RECIST and pathologic responses towards prediction of OS and PFS (in fact more so for SUVmax than TLG in the case of OS). There were, however, two key differences between the studies by Shady et al. and Lastoria et al.: (i) The former study involved radioembolization therapy while the latter involved chemo + anti-angiogenic therapy; (b) the cut-off threshold of response in the former study was set to 30% decrease while for the more conservative latter study, the cut-off was set to 50% decrease in values of PET-based metrics.

In a study by Tam et al. [37] of 70 patients with colorectal liver metastases undergoing different therapies, SUVmean, SUVmax, TLG and MTV were all considered. The measures were not found to be predictive of OS (unlike other studies), but were significant for PFS (HR = 2.46, 2.76, 2.94 and 3.01 for SUVmean, SUVmax, TLG and MTV, respectively). For each measure, threshold optimization was performed using receiver operating characteristic (ROC) analysis. Nonetheless, such ‘optimum cut-off approach’ has the associated problem [38,39] that, even though it amplifies performance in the evaluation set, the probability of false discovery (erroneously obtaining a statistically significant result) increases. This is the reason we did not pursue this approach for further optimization, and instead used median thresholds (values summarized in Table 1 footnote). In addition, our analyses included a range of heterogeneity features (as elaborated in the supplement). We also performed multivariate analysis in which all metrics were comprehensively considered and only those adding value to prediction were selected. By contrast, in the above work, Cox analyses were performed separately for the above-mentioned imaging metrics, and thus their complementary value (if any) could not be deduced. Finally, the present work utilizes methods to account for multiple testing and to discourage overfitting.

In a very recent study by van Helden et al. [40], the authors performed radiomics analysis on pre-treatment PET images of 99 patients with metastatic colorectal cancer undergoing palliative systematic treatment. They found higher volumetric measures (MTV and TLG), asphericity as well as tumor heterogeneity to be predictive of impaired benefit and survival (OS, PFS) following treatment. Though the analysis (primary tumors and metastases) was different from our work (intrahepatic-only tumors), some similar overall trends were observed in that volumetric and few other radiomic features depicted greater predictive value than SUV measures.

Overall, in the present effort we have shown that the number of liver mets, MTV and TLG (as observed or quantified in PET images) were powerful predictors of outcome. Furthermore, multivariate prognostic modeling incorporating radiomic features resulted in improved predictions of outcome. We also evaluated whether classification of metastases into synchronous vs. early metachronous vs. late metachronous improved prediction of outcome, as suggested elsewhere [3]. Furthermore, we assessed in the synchronous cases, whether there was value associated with our knowledge of whether the specific tumors in the analyzed PET images were originally present vs. absent at diagnosis. We did not find these categorizations to be predictive of outcome in OS, PFS or EFS analyses.

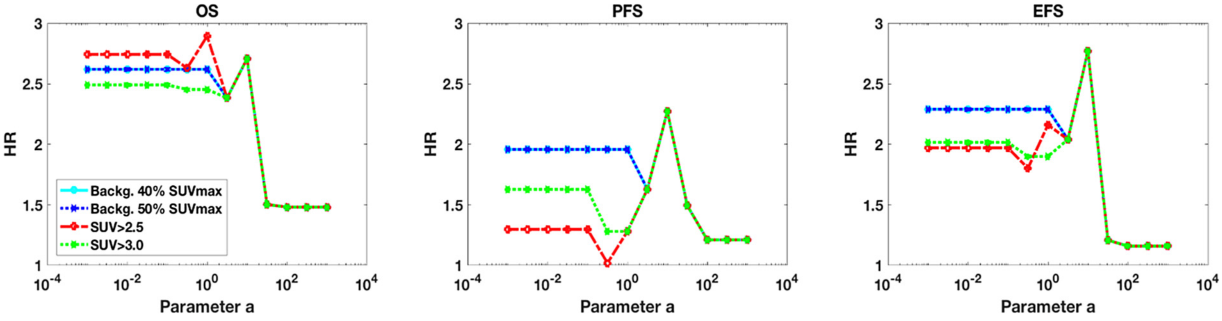

Fig. 5. Plots of HR against varying parameter a in the gETU measure. The plots are shown for (*left*) OS, (*middle*) PFS, and (*right*) EFS, including the four different segmentations. Median thresholding of each metric was utilized for survival analysis. Decreasing parameter a values emphasize volumetric information, while increasing a values emphasize PET uptake intensity.

4.2. Generalized effective total uptake (gETU)

The recently introduced gETU metric [23] (as defined in supplement A and used in our analyses) enables generation of measures (via a free parameter a) that place varying emphases on PET uptake intensity vs. volumetric information, depending on the a value. As $a \to 0$, gETU increasingly emphasizes the volumetric information, becoming equivalent to MTV when $a < < 1$. For $a = 1$, gETU is TLG, equally emphasizing volumetric and intensity information. For $a > 1$, intensity is emphasized, such that for $a > > 1$, gETU becomes equivalent to SUVmax, neglecting volumetric information altogether.

In Fig. 5, we depict OS, PFS and EFS performance by varying parameter a in the gETU metric. It is seen that survival HR performance is especially improved in PFS and EFS when utilizing background-corrected SUVmax thresholding (40% or 50% thresholding methods perform the same for these metrics), outperforming absolute thresholding methods (SUV > 2.5 and SUV > 3.0). Furthermore, an overall trend is seen for thresholding methods that as one shifts towards volumetric information (lower a values), better performance is obtained in the analyses (OS, PFS, and EFS). This indicates the importance of utilizing volumetric information for prediction of outcome from baseline PET images of liver metastases, relative to relying purely on intensity information (e.g. SUVmax which is obtained on the far right). Finally, we saw in our multivariate analyses (results section) that MTV (which corresponds to gETU with $a < < 1$) was retained in the OS and EFS models while SUVmax/peak/mean were not retained in any of the models. Overall, the implication of this plot is that PET-based metabolic tumor volume information is more important than pure uptake information for prediction of outcome in these patients. This is consistent with increasing evidence (as mentioned in the introduction) that, in a range of cancers, volumetric measures can outperform their SUV counterparts for assessment of disease.

4.3. Impact of different statistical criteria and different combinations of metrics

We used the AIC as criterion for multivariate model selection in stepwise Cox regression, accepting a model with an additional parameter if LOGL increased by > 1. More conservative criteria may be considered to further discourage overfitting. This includes use of Wilks' theorem [41] which states that 2(LOGL(model$_2$)-LOGL(model$_1$)) is approximately a chi-squared distribution with degree-of-freedom df = df (model$_2$)-df(model$_1$); setting p-value = 0.05 for accepting a new model$_2$ with an additional parameter (degree of freedom), the new LOGL must be higher by > 1.92. This turns out to be nearly in par here with the Bayesian Information Criterion (BIC), requiring increase in LOGL (for n = 52 patients) by > log(52)/2 = 1.98. It is also close to a required increase in LOGL by > 2 as suggested for the effective parsimony information criterion (EPIC), corresponding to a likelihood ratio test at level 0.05 in the case of testing for the use of one additional parameter between two models [42]. When using the above-mentioned 1.92 threshold (instead of 1 based on AIC), only the first two metrics in Table 1 were retained in the case of OS (number of liver mets and liver-therapy-3mon-prior), with final multivariate HR = 3.77. PFS prediction remained the same (HR = 4.02). In the case of EFS, only the number of liver mets was retained with HR = 2.42 similar to the univariate model.

In our multivariate approach, at each step we accept that metric into the model which increased LOGL the most (and passed the statistical criterion). Nonetheless, it is possible to try different combinations of metrics, to select the combination that at the end produces the largest LOGL. This, however, requires a very large search space and is beyond the scope of our work. We did try one variation: we initialized the multivariate models twice, once by the best performing univariate model, and once by a conventional metric that performed the best (i.e. excluding volumetric or heterogeneity features at first iteration, but then allowing them in subsequent model selection iterations). This was followed by addition of metrics at every iteration that most increased LOGL. Interestingly, the latter initialization resulted in improved performance in one instance (for PFS) which is the result we report in Table 2.

4.4. Considerations and limitations

Our analyzed PET studies were performed in years 2005–2010, all involving 2D-OSEM image reconstruction and 8-mm post-reconstruction Gaussian filtering. Even though more advanced reconstructions (3D-OSEM and PSF modeling [43]) became available in later years, for consistency we only included 2D-OSEM reconstructions which were the only options available for earlier studies. For comparison purposes, we note that images with improved spatial resolution could lead to distinct (probably smaller) VOI volumes than obtained in our work, and for such images, one might need to lower the thresholds to obtain similar VOIs as we do.

We utilized a derivation set for univariate and multivariate analysis. To conclusively establish the proposed models, a distinct validation set is also required. In fact, there is an important frontier, awaiting to be more thoroughly explored in radiomics research, of validating previously derived measures and models. At the same time, to address the issue with false discovery in the context of multiple testing [38], the Benjamini–Hochberg (BH) step-up procedure was utilized for statistical analysis. Furthermore, our multivariate analyses invoked statistical criteria for the acceptance of new metrics in order to discourage overfitting. Not using any such criteria resulted in a larger number of metrics accepted into the multivariate models, with the appearance of improved performance. Nonetheless, we reported the more moderate results that included statistical acceptance criteria.

Our studies group of patients is somewhat heterogeneous in terms of treatment, but we incorporated pre- and post-treatment information as individual features within our predictive modeling to account for this heterogeneity. One may argue that our real-life clinical data set, and the significant findings for it, can render the findings more applicable to a general population of patients with liver mets than results from a very select group. A more select group, at the same time, may result in more significant findings.

Finally, we note that in the present work, the radiomic features utilized were those that could be computed from histograms of segmented PET regions, consistent with existing readily available capabilities of imaging vendor platforms to produce histograms. An exception was the computation of SUVpeak which is available in routine practice. Future work and efforts include more sophisticated recording and analysis of segmented tumors, utilizing the various spatial uptake patterns available in the original images for the computation and analyses of larger sets of radiomic features [26,44,45]. It remains to be seen how effective the above-mentioned histogram-based features are in comparison to broader set of radiomic features.

5. Conclusion

The present work shows that conventional, commonly-employed SUV metrics (SUVmax, SUVpeak, SUVmean) perform relatively poorly in outcome prediction tasks (OS, PFS, EFS) when assessing colorectal liver metastases from FDG PET images. By contrast, use of the number of liver metastasis provided significant performance. This was also the case for volumetric MTV and TLG measures. Furthermore, use of multivariate prognostic modeling while including radiomic features further improved outcome prediction. Our overall finding is that volumetric features outperform SUV-based metrics in the task of clinical outcome prediction, and that prediction can be further enhanced via multivariate models that include volumetric and/or heterogeneity measures. This improved prediction of clinical outcome has the potential to be used for non-invasive selection of patients for individual treatment modality or participation in clinical trials of different treatment regimes.

Author Contributions

AR: Led the overall project in terms of data analysis and writing the manuscript.

KPBF: PET image analyses, collecting non-PET data, interpretation of results, contributed to the final manuscript.

SA: Developed standardized metrics (radiomics) and pipeline used in data analysis.

LL: Contributed to statistical modeling and analysis of data.

CRS: Discussion of data and interpretation of results, and contributed to the final manuscript.

RMS: Clinical motivation for the work, discussion of data and interpretation of results.

AM: Discussion of data and interpretation of results, and contributed to the final manuscript.

SK: PET scanning of liver patients, PET image analyses, Collecting non-PET data, contributed to the final manuscript.

JH: PET image analyses, Collecting non-PET clinical data.

OLM: Data preprocessing and validation, histogram generation, interpretation of results, contributed to the final manuscript.

Financial disclosure

Dr. Keiding: Danish Council for Independent Research (Medical Sciences, 4004-00022).

Dr. Schmidtlein: MSK Cancer Center Support Grant/Core Grant (P30 CA008748).

Dr. Lu: National Natural Science Foundation of China. No disclosures exist for any other authors.

IRB Statement

Formal approval for access and analysis of patient data was obtained from the Danish Patient Safety Authority and the study was also approved by the Danish Data Protection Agency.

Acknowledgements

This work was supported by the Danish Council for Independent Research (Medical Sciences, 4004-00022), the MSK Cancer Center Support Grant/Core Grant (P30 CA008748), and the National Natural Science Foundation of China under grants 61628105. We also acknowledge very helpful discussions with Dr. Ciprian Crainiceanu.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ejrad.2019.02.006.

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int. J. Cancer 136 (5) (2015).
[2] S. Manfredi, C. Lepage, C. Hatem, O. Coatmeur, J. Faivre, A.-M. Bouvier, Epidemiology and management of liver metastases from colorectal cancer, Ann. Surg. 244 (2) (2006) 254.
[3] R. Adam, A. de Gramont, J. Figueras, N. Kokudo, F. Kunstlinger, E. Loyer, G. Poston, P. Rougier, L. Rubbia-Brandt, A. Sobrero, Managing synchronous liver metastases from colorectal cancer: a multidisciplinary international consensus, Cancer Treat. Rev. 41 (9) (2015) 729–741.
[4] R.L. Wahl, Principles and Practice of PET and PET/CT, 2nd ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2008.
[5] R.L. Wahl, H. Jacene, Y. Kasamon, M.A. Lodge, From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors, J. Nucl. Med. 50 (Suppl. 1) (2009) 122S–150S.
[6] M.A. Lodge, M.A. Chaudhry, R.L. Wahl, Noise considerations for PET quantification using maximum and peak standardized uptake value, J. Nucl. Med. 53 (7) (2012) 1041–1047.
[7] E.H. Dibble, A.C.L. Alvarez, M.-T. Truong, G. Mercier, E.F. Cook, R.M. Subramaniam, 18F-FDG metabolic tumor volume and total glycolytic activity of oral cavity and oropharyngeal squamous cell cancer: adding value to clinical staging, J. Nucl. Med. 53 (5) (2012) 709–715.
[8] J. Davison, G. Mercier, G. Russo, R.M. Subramaniam, PET-based primary tumor volumetric parameters and survival of patients with non—small cell lung carcinoma, Am J Roentgenol 200 (3) (2013) 635–640.
[9] K. Pak, G.J. Cheon, H.Y. Nam, S.J. Kim, K.W. Kang, J.K. Chung, E.E. Kim, D.S. Lee, Prognostic value of metabolic tumor volume and total lesion glycolysis in head and neck cancer: a systematic review and meta-analysis, J. Nucl. Med. 55 (6) (2014) 884–890.
[10] I.S. Ryu, J.S. Kim, J.L. Roh, K.J. Cho, S.H. Choi, S.Y. Nam, S.Y. Kim, Prognostic significance of preoperative metabolic tumour volume and total lesion glycolysis measured by (18)F-FDG PET/CT in squamous cell carcinoma of the oral cavity, Eur. J. Nucl. Med. Mol. Imaging 41 (3) (2014) 452–461.
[11] G.C. Park, J.S. Kim, J.L. Roh, S.H. Choi, S.Y. Nam, S.Y. Kim, Prognostic value of metabolic tumor volume measured by 18F-FDG PET/CT in advanced-stage squamous cell carcinoma of the larynx and hypopharynx, Ann. Oncol. 24 (1) (2013) 208–214.
[12] S.J. Lee, J.Y. Choi, H.J. Lee, C.H. Baek, Y.I. Son, S.H. Hyun, S.H. Moon, B.T. Kim, Prognostic value of volume-based (18)F-fluorodeoxyglucose PET/CT parameters in patients with clinically node-negative oral tongue squamous cell carcinoma, Korean J. Radiol. 13 (6) (2012) 752–759.
[13] M. Taghipour, R. Wray, S. Sheikhbahaei, J.L. Wright, R.M. Subramaniam, FDG avidity and tumor burden: survival outcomes for patients with recurrent breast cancer, Am. J. Roentgenol. 206 (4) (2016) 846–855.
[14] C. Marcus, R. Wray, M. Taghipour, W. Marashdeh, S.J. Ahn, E. Mena, R.M. Subramaniam, JOURNAL CLUB: Value of quantitative FDG PET/CT volumetric biomarkers in recurrent colorectal Cancer patient survival, Am. J. Roentgenol. 207 (2) (2016) 257–265.
[15] J.M.M. Rogasch, P. Hundsdoerfer, F. Hofheinz, F. Wedel, I. Schatka, H. Amthauer, C. Furth, Pretherapeutic FDG-PET total metabolic tumor volume predicts response to induction therapy in pediatric Hodgkin's lymphoma, BMC Cancer 18 (1) (2018) 521.
[16] D. Albano, M. Bertoli, M. Battistotti, C. Rodella, M. Statuto, R. Giubbini, F. Bertagna, Prognostic role of pretreatment 18F-FDG PET/CT in primary brain lymphoma, Ann. Nucl. Med. 32 (8) (2018) 532–541.
[17] V. Kumar, Y.H. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H.J.W.L. Aerts, A. Dekker, D. Fenstermacher, D.B. Goldgof, L.O. Hall, P. Lambin, Y. Balagurunathan, R.A. Gatenby, R.J. Gillies, Radiomics: the process and the

challenges, Magn. Reson. Imaging 30 (9) (2012) 1234–1248.

[18] M.C. Asselin, J.P.B. O'Connor, R. Boellaard, N.A. Thacker, A. Jackson, Quantifying heterogeneity in human tumours using MRI and PET, Eur. J. Cancer 48 (4) (2012) 447–455.

[19] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G.P.M. van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J.W.L. Aerts, Q.-C. Consortium, Radiomics: Extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (4) (2012) 441–446.

[20] S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, G.J.R. Cook, Quantifying tumour heterogeneity in F-18-FDG PET/CT imaging by texture analysis, Eur. J. Nucl. Med. Mol. I 40 (1) (2013) 133–140.

[21] M. Hatt, F. Tixier, L. Pierce, P.E. Kinahan, C.C. Le Rest, D. Visvikis, Characterization of PET/CT images using texture analysis: the past, the presenta… any future? Eur. J. Nucl. Med. Mol. I 44 (1) (2017) 151–165.

[22] R. Boellaard, M.J. O'Doherty, W.A. Weber, F.M. Mottaghy, M.N. Lonsdale, S.G. Stroobants, W.J. Oyen, J. Kotzerke, O.S. Hoekstra, J. Pruim, FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0, Eur. J. Nucl. Med. Mol. I 37 (1) (2010) 181.

[23] A. Rahmim, C.R. Schmidtlein, A. Jackson, S. Sheikhbahaei, C. Marcus, S. Ashrafinia, M. Soltani, R.M. Subramaniam, A novel metric for quantification of homogeneous and heterogeneous tumors in PET for enhanced clinical outcome prediction, Phys. Med. Biol. 61 (2016) 227–242.

[24] I. El Naqa, P.W. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, W.L. Thorstad, J.O. Deasy, Exploring feature-based approaches in PET images for predicting cancer treatment outcomes, Pattern Recognit. 42 (6) (2009) 1162–1171.

[25] F.H.P. van Velden, P. Cheebsumon, M. Yaqub, E.F. Smit, O.S. Hoekstra, A.A. Lammertsma, R. Boellaard, Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies, Eur. J. Nucl. Med. Mol. I 38 (9) (2011) 1636–1647.

[26] A. Zwanenburg, S. Leger, M. Vallières, S. Löck, Image Biomarker Standardisation Initiative - Feature Definitions, CoRR abs/1612.07003 (2016).

[27] S. Lo, I. Li, T. Tsou, L. See, Non-significant in univariate but significant in multivariate analysis: a discussion with examples, Changgeng Yi Xue Za Zhi 18 (2) (1995) 95–101.

[28] L.F. de Geus-Oei, B. Wiering, P.F. Krabbe, T.J. Ruers, C.J. Punt, W.J. Oyen, FDG-PET for prediction of survival of patients with metastatic colorectal carcinoma, Ann. Oncol. 17 (11) (2006) 1650–1655.

[29] E.G. Dimitrova, B.G. Chaushev, N.V. Conev, J.K. Kashlov, A.K. Zlatarov, D.P. Petrov, H.B. Popov, N.T. Stefanova, A.D. Klisarova, K.Z. Bratoeva, Role of the pretreatment 18F-fluorodeoxyglucose positron emission tomography maximal standardized uptake value in predicting outcomes of colon liver metastases and that value's association with Beclin-1 expression, Biosci. Trends 11 (2) (2017) 221–228.

[30] S. De Bruyne, N. Van Damme, P. Smeets, L. Ferdinande, W. Ceelen, J. Mertens, C. Van de Wiele, R. Troisi, L. Libbrecht, S. Laurent, Value of DCE-MRI and FDG-PET/CT in the prediction of response to preoperative chemotherapy with bevacizumab for colorectal liver metastases, Br. J. Cancer 106 (12) (2012) 1926.

[31] D. Vriens, H.W. Van Laarhoven, J.J. van Asten, P.F. Krabbe, E.P. Visser, A. Heerschap, C.J. Punt, L.-F. de Geus-Oei, W.J. Oyen, Chemotherapy response monitoring of colorectal liver metastases by dynamic Gd-DTPA–enhanced MRI perfusion parameters and 18F-FDG PET metabolic rate, J. Nucl. Med. 50 (11) (2009) 1777–1784.

[32] F.A. Kotasidis, C. Tsoumpas, A. Rahmim, Advanced kinetic modelling strategies: towards adoption in clinical PET imaging, Clin. Transl. Imaging 2 (3) (2014) 219–237.

[33] N.A. Karakatsanis, M.A. Lodge, A.K. Tahari, Y. Zhou, R.L. Wahl, A. Rahmim, Dynamic whole body PET parametric imaging: I. Concept, acquisition protocol optimization and clinical application, Phys. Med. Biol. 58 (20) (2013) 7391–7418.

[34] S.A. Gulec, R.R. Suthar, T.C. Barot, K. Pennington, The prognostic value of functional tumor volume and total lesion glycolysis in patients with colorectal cancer liver metastases undergoing 90 Y selective internal radiation therapy plus chemotherapy, Eur. J. Nucl. Med. Mol. I 38 (7) (2011) 1289–1295.

[35] W. Shady, S. Kishore, S. Gavane, R.K. Do, J.R. Osborne, G.A. Ulaner, M. Gonen, E. Ziv, F.E. Boas, C.T. Sofocleous, Metabolic tumor volume and total lesion glycolysis on FDG-PET/CT can predict overall survival after 90Y radioembolization of colorectal liver metastases: A comparison with SUVmax, SUVpeak, and RECIST 1.0, Eur. J. Radiol. 85 (6) (2016) 1224–1231.

[36] S. Lastoria, M.C. Piccirillo, C. Caracò, G. Nasti, L. Aloj, C. Arrichiello, E.D.L. Di Castelguidone, F. Tatangelo, A. Ottaiano, R.V. Iaffaioli, Early PET/CT scan is more effective than RECIST in predicting outcome of patients with liver metastases from colorectal cancer treated with preoperative chemotherapy plus bevacizumab, J. Nucl. Med. 54 (12) (2013) 2062–2069.

[37] H.H. Tam, G.J. Cook, I. Chau, B. Drake, I. Zerizer, Y. Du, D. Cunningham, D.-M. Koh, S.S. Chua, The role of routine clinical pretreatment 18F-FDG PET/CT in predicting outcome of colorectal liver metastasis, Clin. Nucl. Med. 40 (5) (2015) e259–e264.

[38] A. Chalkidou, M.J. O'Doherty, P.K. Marsden, False discovery rates in PET and CT studies with texture features: a systematic review, PLoS One 10 (5) (2015).

[39] S.G. Hilsenbeck, G.M. Clark, W.L. McGuire, Why do so many prognostic factors fail to pan out? Breast Cancer Res. Treat. 22 (3) (1992) 197–206.

[40] E.J. van Helden, Y.J.L. Vacher, W.N. van Wieringen, F.H.P. van Velden, H.M.W. Verheul, O.S. Hoekstra, R. Boellaard, C.W. Menke-van der Houven van Oordt, Radiomics analysis of pre-treatment [(18)F]FDG PET/CT for patients with metastatic colorectal cancer undergoing palliative systemic treatment, Eur. J. Nucl. Med. Mol. Imaging 45 (13) (2018) 2307–2317.

[41] S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. 9 (1938) 60–62.

[42] R.T. Shinohara, C.M. Crainiceanu, B.S. Caffo, D.S. Reich, Longitudinal Analysis of Spatiotemporal Processes: A Case Study of Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Multiple Sclerosis, Johns Hopkins University, Dept. of Biostatistics Working Papers Working Paper 231, 2011.

[43] A. Rahmim, J. Qi, V. Sossi, Resolution modeling in PET imaging: theory, practice, benefits, and pitfalls, Med. Phys. 40 (6) (2013) 064301.

[44] R.T.H. Leijenaar, S. Carvalho, E.R. Velazquez, W.J.C. Van Elmpt, C. Parmar, O.S. Hoekstra, C.J. Hoekstra, R. Boellaard, A.L.A.J. Dekker, R.J. Gillies, H.J.W.L. Aerts, P. Lambin, Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability, Acta Oncol. 52 (7) (2013) 1391–1397.

[45] W. Lv, Q. Yuan, Q. Wang, J. Ma, J. Jiang, W. Yang, Q. Feng, W. Chen, A. Rahmim, L. Lu, Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT, Eur. Radiol. 8 (2018) 3245–3254.