Nuclear Medicine Communications

Revalidation of PET/CT criteria (Hopkins criteria) for the assessment of therapeutic response in lung cancer patients: Inter-reader reliability, accuracy and survival outcomes

--Manuscript Draft--

Manuscript Number:	NMC-11-3252R1
Full Title:	Revalidation of PET/CT criteria (Hopkins criteria) for the assessment of therapeutic response in lung cancer patients: Inter-reader reliability, accuracy and survival outcomes
Article Type:	Original Study
Keywords:	Lung cancer; PET/CT; response assessment; survival.
Corresponding Author:	Francesco FRAIOLI, M.D. University College London Institute of Nuclear Medicine UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University College London Institute of Nuclear Medicine
Corresponding Author's Secondary Institution:	
First Author:	Khulood Al Riyami
First Author Secondary Information:	
Order of Authors:	Khulood Al Riyami
	Noor Al Nuaimi
	Ruta Kliokyte
	Stefan Voo
	Andrew Thornton
	Jamshed Bomanji
	Francesco FRAIOLI, M.D.
Order of Authors Secondary Information:	
Manuscript Region of Origin:	UNITED KINGDOM
Abstract:	Background/Aim Systematic reporting using qualitative evaluation of PET/CT results has been demonstrated to be very accurate and reproducible in post-therapy assessment of lung cancer (so-called Hopkins criteria). Our aim was to test, in a different cohort of patients, the Hopkins criteria for assessment of therapeutic response in lung cancer and to compare the results with those obtained using a semi-quantitative evaluation of uptake.Methods This is a retrospective study. A total of 85 patients with known lung cancer who underwent ^{18}F-FDG PET/CT assessment within 24 weeks (mean 7.9 weeks) of completion of treatment were included. Treatments included surgical resection, chemotherapy, radiation therapy, immunotherapy or combinations thereof. PET/CT interpretation was done by two nuclear medicine physicians, and discrepancies were resolved by a third interpreter. Studies were scored both according to the Hopkins criteria using qualitative assessment of tracer uptake for the primary tumour, loco-regional disease in the mediastinum and distant metastatic sites and by applying the same 5-point score using a semi-quantitative measure, SUV_{max}. Overall scores of 1, 2 and 3 were considered negative for residual disease, while scores of 4 and 5 were

considered positive. Patients were followed up for a median of 18.5 months (range 2–139 months). Kaplan-Meier plots with a Mantel-Cox log-rank test were performed, considering death as the endpoint. Inter-reader variability was assessed using percent agreement and kappa statistics.Results

The Cohen κ coefficient analysis showed substantial agreement between the two interpreters on the five-point Hopkins criteria scoring, with a κ of 0.73. There was almost perfect agreement between the interpreters with respect to classification as positive or negative according to the Hopkins criteria, with a κ of 0.89. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the Hopkins criteria were 88.5% (95%CI 80.6%–96.5%), 79.2% (95%CI 63.2%–95.1%), 91.5% (95%CI 84.4%–98.6%), 73.1% (95%CI 61.8%–84.4%) and 85.9% (95%CI 78.5%–93.3%) respectively. There was almost perfect agreement between the qualitative and semi-quantitative scoring with a κ of 0.87, with sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the semi-quantitative Hopkin’s criteria of 86.9% (95% CI 78.4%–95.4%), 79.2% (95%CI 62.9%–95.4%), 91.4% (95%CI 84.2%–98.6%), 70.4% (95%CI 58.6%–82.1%), and 84.7% (95%CI 80.8%–92.4%) respectively.Conclusion

The use of Hopkins criteria for post-therapy assessment in patients with lung cancer represents an easy and reproducible method with substantial to almost perfect inter-observer agreement and high positive predictive value and accuracy; moreover, it is easily understood by referring physicians. Additionally, there was no significant difference when applying a semi-quantitative measure to the same 5 point score.

Title: Revalidation of PET/CT criteria (Hopkins criteria) for the assessment of therapeutic response in lung cancer patients: Inter-reader reliability, accuracy and survival outcomes

Short title: Revalidation of Hopkins criteria in lung cancer

Authors: Khulood Al Riyami[1,2], Noor Al Nuaimi[1], Ruta Kliokyte[3], Andrew Thornton[1], Jamshed Bomanji[1] and Francesco Fraioli[1]

Affiliation: 1. Institute of nuclear medicine, University College Hospital London, UK

2.Department of radiology and molecular imaging, Sultan Qaboos University Hospital, Oman

3.Centre of Radiology and Nuclear Medicine, Vilnius University Santaros clinics, Lithuania

Correspondence to: Francesco Fraioli

Number or words:	4210
Number of tables:	3
Number of figures:	2

Abstract

Background/Aim. Systematic reporting using qualitative evaluation of PET/CT results has been demonstrated to be very accurate and reproducible in post-therapy assessment of lung cancer (so-called Hopkins criteria). Our aim was to test, in a different cohort of patients, the Hopkins criteria for assessment of therapeutic response in lung cancer and to compare the results with those obtained using a semi-quantitative evaluation of uptake.

Methods. This is a retrospective study. A total of 85 patients with known lung cancer who underwent ^{18}F-FDG PET/CT assessment within 24 weeks (mean 7.9 weeks) of completion of treatment were included. Treatments included surgical resection, chemotherapy, radiation therapy, immunotherapy or combinations thereof. PET/CT interpretation was done by two nuclear medicine physicians, and discrepancies were resolved by a third interpreter. Studies were scored both according to the Hopkins criteria using qualitative assessment of tracer uptake for the primary tumour, loco-regional disease in the mediastinum and distant metastatic sites and by applying the same 5-point score using a semi-quantitative measure, SUV_{max}. Overall scores of 1, 2 and 3 were considered negative for residual disease, while scores of 4 and 5 were considered positive. Patients were followed up for a median of 18.5 months (range 2–139 months). Kaplan-Meier plots with a Mantel-Cox log-rank test were performed, considering death as the endpoint. Inter-reader variability was assessed using percent agreement and kappa statistics.

Results. The Cohen κ coefficient analysis showed substantial agreement between the two interpreters on the five-point Hopkins criteria scoring, with a κ of 0.73. There was almost perfect agreement between the interpreters with respect to classification as positive or negative according to the Hopkins criteria, with a κ of 0.89. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the Hopkins criteria were 88.5% (95%CI 80.6%–96.5%), 79.2% (95%CI 63.2%–95.1%), 91.5% (95%CI 84.4%–98.6%), 73.1% (95%CI 61.8%–84.4%) and 85.9% (95%CI 78.5%–93.3%) respectively. There was almost perfect agreement between the qualitative and semi-quantitative scoring with a κ of 0.87, with sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the semi-quantitative Hopkin's criteria of 86.9% (95% CI 78.4%–95.4%), 79.2% (95%CI 62.9%–95.4%), 91.4% (95%CI 84.2%–98.6%), 70.4% (95%CI 58.6%–82.1%), and 84.7% (95%CI 80.8%–92.4%) respectively.

Conclusion. The use of Hopkins criteria for post-therapy assessment in patients with lung cancer represents an easy and reproducible method with substantial to almost perfect inter-observer agreement and high positive predictive value and accuracy; moreover, it is easily understood by referring physicians. Additionally, there was no significant difference when applying a semi-quantitative measure to the same 5 point score.

Key Words. PET/CT; lung cancer; therapy assessment; Hopkins criteria

Introduction

Lung cancer is the leading cause of cancer death among men and the second leading cause of cancer death among women worldwide, with over 2 million new cases in 2018 [1]. In the United states, the estimates for lung cancer for 2019 is over 200,000 new cases and over 140,000 deaths from lung cancer [2]. Similarly, in the United Kingdom, it was the most common cause of cancer death in 2014 [3]. Despite the advances in the available therapeutic options, recurrences after lung cancer surgery typically happen rapidly, with 90%–95% occurring within 5 years [4]. Survival time following local or distant recurrence averages less than a year, including among patients who receive salvage treatment [5]. In this context, the assessment of therapeutic response could change the clinical management of the patient and potentially improve survival [6].

Fluorine-18 fluorodeoxyglucose (FDG) PET/CT has been established as an important imaging method for the diagnostic work-up of lung cancer patients [7, 8], though the need for an established systematic and reproducible interpretation system has also been recognised [9]. Evaluation of response using PET can be performed visually or semi quantitatively through comparison of a lesion's SUV max with the background liver uptake and the mediastinal blood pool.

The recently introduced Hopkins criteria for lung cancer response assessment in PET/CT seems to offer substantial inter-observer agreement as well as high sensitivity and specificity to predict survival in lung cancer patients, irrespective of tumour histology and treatment [6]. The Hopkins criteria does not suggest comparison using the SUV_{max}; although there are other standards such as PERCIST, those are of not easily adopted in clinical practice, hence we will consider the semi-quantitative extension to the Hopkin's criteria where assessment of uptake will be compared against the SUV_{max}.

The aim of this study was to externally validate the reproducibility of the Hopkins criteria and compare their use with a semi-quantitative method of evaluation of uptake.

Materials and Methods

This is a retrospective, single-centre cohort diagnostic study with patient inclusion conforming to the principles outlined in the Helsinki Declaration II. Due to the retrospective character of the study, ethical approval was waived by the institutional ethics committee.

Eligible Patients and Follow-up

A retrospective evaluation was conducted on consecutive patients with lung cancer who had undergone PET/CT scans for post-therapy assessment.

Inclusion criteria. The study included patients with biopsy-proven lung cancer who underwent post-therapy ^{18}F-FDG PET/CT at our institute between 2007 and 2015 for the evaluation of treatment response. All patients were treated with surgical resection, chemotherapy, radiotherapy or immunotherapy, or a combination of these treatments, and then underwent ^{18}F-FDG PET/CT within 24 weeks of therapy.

Patient demographics, clinical history and clinical data were collected from the electronic medical records for up to 6 months following PET/CT.

Exclusion criteria. Patients with no available follow-up, patients with PET/CT imaging >24 weeks post therapy, patients with no available imaging and those with concurrent malignancies were excluded.

Image Analysis

Hopkins Criteria for Post-Therapy Assessment on PET/CT.

Studies were scored using a qualitative five-point scale, for the primary tumour, locoregional disease in the mediastinum and distant nodal or metastatic sites (Table 1*)*. The visual activity in the mediastinal blood pool and in the liver was taken as the background blood pool for reference.

Semi-quantitative Criteria for Post-therapy Assessment on PET/CT

The same five-point scale was applied using SUV_{max} as a semi-quantitative measure of tracer. The mediastinal blood pool, liver background and SUV_{max} within the tumour, nodal or distant disease were recorded and categorised as above.

Definition of Positive and Negative PET/CT Studies. On the basis of the qualitative five-point scale, the studies were grouped as positive or negative for the primary tumour, locoregional disease in the mediastinum and distant metastatic lesions. Overall assessment was denoted by the overall score, which was the highest score among the scores for the primary tumour and locoregional and distant metastatic lesions, if present. Scores 1, 2 and 3 were considered negative for residual tumour and scores 4 and 5 were considered positive.

All PET/CT studies were retrieved from the institutional Picture Archiving and Communication System (PACS) and reviewed on a Carestream Vue PACS workstation/viewing platform (version 12.1.5.7014, Carestream Health Inc). All images were interpreted by two reviewers independently (readers 1 and 2). Reader 1 was a board-certified radiologist with more than 4 years' subspecialty training in nuclear medicine, and reader 2 was a board-certified medical physician with subspecialty in nuclear medicine and expertise in lung oncologic imaging. In order to reach a consensus, any discrepancies were resolved by a third interpreter, a senior consultant dual board-certified in nuclear medicine and radiology who was the main nuclear medicine representative at the local lung cancer multidisciplinary meeting (MDM). This final consensus score was used to determine the final Hopkins score.

Outcome measures

As in the original article on use of the Hopkins criteria (6), histological confirmation of PET/CT-positive lesions, alternative imaging modalities or clinical follow-up of 6 months after PET/CT were considered as the reference standard.

The sensitivity and specificity, positive predictive value, negative predictive value and accuracy of the post-therapy PET/CT assessment criteria, along with 95% confidence intervals (CIs), were calculated by constructing a 2×2 contingency table (cross-relating PET/CT results of the reference standards).

Overall survival for all cases was defined as the time interval in months between the date of the post-therapy PET/CT study and the date of death or loss to follow-up. The date of the study was recorded from the radiology information system (RIS) and the date of death was extracted from the electronic medical records.

Statistical analysis

Descriptive values are presented as mean (with standard deviation) or median (with 25th to 75th percentile range) if the data were not normally distributed. Categorical variables are presented as frequency (with percentage). The Cohen κ co-efficient was calculated to measure inter-interpreter agreement and inter-criteria agreement. Survival probabilities were generated using Kaplan-Meier survival curves and compared using the Mantel-Cox log-rank test. Univariate and multivariate Cox regression analyses were performed, considering death as the endpoint. Subgroup analysis was performed to assess the impact of histological subtype and prior treatment on the prognostic value of the Hopkins score. The statistical significance level was set at a p value of less than 0.05. Statistical analysis was performed using R 3.3.1.

Results

Patient Characteristics and Follow-up

Eighty-five patients were included in the study (45 male, 40 female; mean age ± SD, 63.5±10.2 years). A history of smoking was present in 55 patients (64.7%). The histological subtype of the primary malignancy was identified as small cell lung cancer (SCLC) in 10 patients (11.8%) and non-small cell lung cancer (NSCLC) in 74 (87.1%); histological subtype was not identified in one patient. The demographic details of the 85 patients included in the study are summarised in Table *2*. The median follow-up of these patients was 18.5 months (range 2–139 months) after completion of the post-treatment assessment PET/CT.

Time Interval to Post-treatment PET/CT

Post-treatment ^{18}F-FDG PET/CT for assessment of therapeutic response was performed between 0 and 24 weeks after completion of treatment. The average interval between the date of completion of treatment and the post-treatment ^{18}F-FDG PET/CT was 7.9 weeks (median 6 weeks, range 0–24 weeks).

Interpreter Classification of PET/CT studies (by Hopkins)

On the basis of the final Hopkins qualitative criteria scores, PET/CT studies were characterised as positive in 59 patients (69.4%) and as negative in 26 (30.6%). On the 59 positive PET/CT studies, the most avid residual disease was identified at the primary site in 27 patients (45.8%), at sites of nodal disease in 23 (39.0%), at sites of distant metastases in 7 (11.9%) and in the pleura in 2 (3.4%). Of the 26 PET/CT studies characterised as negative, 14 (53.6%) were scored as 1 or 2 and 12 (46.2%) as 3.

The accuracy of the scoring system was assessed by imaging in 66 cases (77.6%), histology in ten (11.8%) and clinical follow-up in nine (10.6%). Table 3 summarises the results of this follow-up. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the scoring system were 88.5% (95%CI 80.6%–96.5%), 79.2% (95%CI 63.2%–95.1%), 91.5% (95%CI 84.4%–98.6%), 73.1% (95%CI 61.8%–84.4%) and 85.9% (95%CI 78.5%–93.3%) respectively.

Interpreter Classification of PET/CT studies (by semi-quantitative Hopkins extension)

On the basis of the final semi-quantitative criteria scores, PET/CT studies were characterised as positive in 58 patients (68.2%) and as negative in 26 patients (30.6%). One patient had to be excluded from assessment as height and weight were not provided and SUV_{max} could not therefore be calculated. On the 58 positive PET/CT studies, the most avid residual disease was identified at the primary site in 27 patients (46.5%), at sites of nodal disease in 22 (37.9%), at sites of distant metastases in 7 (12.1%) and in the pleura in 2 (3.4%). Of the 26 PET/CT studies characterised as negative, 14 (53.6%) were scored as 1 or 2 and 12 (46.2%) as 3.

The accuracy of the scoring system was assessed by imaging in 66 cases (78.6%), histology in nine (10.7%) and clinical follow-up in nine (10.7%). The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the semi-quantitative Hopkin's criteria were 86.9% (95% CI 78.4%–95.4%), 79.2% (95%CI 62.9%–95.4%), 91.4% (95%CI 84.2%–98.6%), 70.4% (95%CI 58.6%–82.1%), and 84.7% (95%CI 80.8%–92.4%) respectively.

Interobserver agreement

The Cohen κ coefficient analysis indicated substantial agreement between the two interpreters (R1 and R2) on the five-point qualitative Hopkins criteria scoring, with a κ of 0.73. Discrepancies between the two interpreters (15 patients, 17.2%) were resolved by the third interpreter (R3). There was almost perfect agreement between R1 and R2 in terms of positive versus negative classification according to the Hopkins criteria, with a κ of 0.89 (discrepancies occurred in only four patients, 4.7%). When scoring was performed using the semi-quantitative measure, SUV_{max}, substantial agreement was again observed between the two interpreters, with a κ of 0.72, but discrepancies occurred in 19 patients (22%), i.e. four more than in the qualitative assessment.

Intercriteria agreement

There was also almost perfect agreement between the final five point qualitative and semi-quantitative scoring with a κ of 0.87 – this increased to κ of 0.97 when scoring was qualified as positive versus negative with only 1 discrepancy (1.2%). 1 patient had to be excluded from the quantitative analysis as SUVs could not be generated due to a lack of patient height and weight information.

Survival Outcome in All Patients

The median follow-up of the study population was 20.5 months (range 2–139 months) after completion of the post-treatment assessment PET/CT and 60 patients died during follow-up (70.6%). The median survival of the Hopkins criteria-positive group was 17.2 months (95% CI 13.7–28.6 months) and 44 (74.6%) patients died in this group. The median survival of the Hopkins criteria-negative group was 32.4 months (95% CI, 24.6–), and 16 (61.5%) patients died in this group. The Kaplan-Meier survival analysis showed a non-significant difference in overall survival (OS) (Score (logrank) test = 2.66 on 1 df, p=0.1027), with a hazard ratio (HR) of 1.61 (95% CI 0.90–2.86) (Figure 1). Hopkins criteria positivity was not statistically significant in multivariate analysis when compared with stage, age or sex. Results were similar for the semi-quantitative scoring system.

In overall assessment using the five-point interpretation scale, there was no significant in difference in OS between patients who scored 1 or 2 (n=15) versus those who scored 3 (n=13) versus those who scored 4 or 5 (n=59), (Score (logrank) test = 3.06 on 2 degree of freedom [df], p=0.2162). There was, however, a significant difference in OS based on Hopkins criteria for positivity, with Score (logrank) test = 11.0 on 4 df, p=0.027 and there was a significant difference in OS if those with Score 5 (n=44) were compared against those scoring less than 5 (n=41), Score (logrank) test = 10.43 on 1 df, p=0.001243 HR 2.4 (95%CI 1.38–4.03). Results were similar for the semi-quantitative scoring system.

Survival Outcomes: Impact of Tumour Histology and Treatment Modality

Based on tumour histology, ten patients (11.8%) were diagnosed with SCLC and 74 (87.1%) with NSCLC (one case was unspecified). The Kaplan-Meier analysis showed a difference between OS in the eight SCLC patients who had a positive PET/CT result (median survival 6.75 months) and in the two who had a negative PET/CT result (median survival 10.61 months), with an HR of 2.83; however, this difference was not statistically significant due to the small numbers of patients involved. In patients with NSCLC (n=74) or an unspecified histology (n=1), OS was shorter in those with a positive PET/CT (n=51, median survival 20.6 months) than in those with a negative PET/CT (n=24, median survival 32.4 months), with an HR of 1.44; again, however, this result was not statistically significant. When restricted to those patients who scored 5, the median survival difference for SCLC was 5.11 months versus 10.9 HR 2.25, again non-significant, but was significant for NSCLC (or unspecified) with 14.7 months versus 33.8, HR 2.18 (95%CI 1.23–3.89) Score (logrank) test = 7.36 on 1 df, p=0.00665. Results were similar for the semi-quantitative scoring system.

There was no significant difference in OS when patients were distinguished according to preceding treatment; however, there was again a trend towards a difference in survival between those with a positive and those with a negative PET/CT by Hopkins Criteria, the difference being greater in those who were treated with radiotherapy or surgery [median OS 16.8 months vs 25.4 months (positive vs negative PET/CT, respectively) for those treated with chemotherapy or immunotherapy alone, compared with 18.9 months vs 45.1 months for those who received radiotherapy or surgery]. Again when restricted to those patients who scored 5 there is a non-significant but almost significant difference in survival between those where treated with chemotherapy or immunotherapy alone (median OS 11.4 months vs 33.8 months), but there is a significant difference in survival with those with score 5 treated with radiotherapy and surgery versus those with scores of 4 or less, (median OS 14.2 vs 32.4 months, HR 2.4 (95%CI 1.14–5.08) Score (logrank) test = 5.69 on 1 df, p=0.01705. Results were similar for the semi-quantitative scoring system although the difference between survival at score 5 vs <5 is significant in those treated with chemotherapy or immunotherapy alone.

Discussion

The role of ^{18}F-FDG PET/CT in treatment response assessment has been widely established in other types of malignancy such as lymphoma, using the Deauville criteria [12, 13].

The use of ^{18}F-FDG PET/CT in lung cancer staging is well established and demonstrated that PET response assessment is much more strongly correlated with survival than response measured by CT scanning [14]; however there is less well established consensus or recommendations on which reporting system is best to use, either visual or quantitative. In our study, we sought to answer two questions, whether the results from visual analysis were reproducible, as previously demonstrated, and second if there is an agreement between the visual and quantitative scoring systems.

In our study we can confirm the results shown by Sheikhbahaei et al. that the Hopkins criteria permit reproducible qualitative assessment of therapeutic response using visual ^{18}F-FDG uptake and can be of a great value for patient care. When using the qualitative five-point Hopkins scoring system, similar to the data presented [6] we observed substantial agreement between the readers and almost perfect agreement when categorizing the patients into positive and negative for disease as per the criteria.

Several studies have demonstrated the added value of post-treatment ^{18}F-FDG PET in the prognostication of patients with lung cancer. These studies reported longer survival in patients with a complete metabolic response, post-therapy reduction in ^{18}F-FDG uptake and changes in total lesion glycolysis and metabolic tumour volume [7, 14–18]. As observed by Sheikhbahaei et al [6], there was a trend towards better overall survival in patients with negative post-treatment scans and in those with scores of 1 and 2; however, in our study these differences were not statistically significant, which may be attributed to the population size.

As for the second purpose, in our study we demonstrated no significant difference in the inter-reader and inter criteria agreement using the qualitative Hopkins criteria and the same five-point scoring system using SUV_{max} as a semi-quantitative measure of tracer uptake. This highlights the fact that the simplified method of using visual assessment for scoring is a reliable technique.

Furthermore, to strengthen the visual assessment criteria it should be recognised that SUV_{max} values can be affected by patient related factors such as fasting blood glucose levels, altered bio-distribution of 18F-FDG which can occur in morbid obesity and technical parameters such as varying uptake time, image noise, partial volume effect and differences in acquisition techniques such as the number of iterations [19, 20].

Finally, although we did not find significant difference in OS when restricted to subgroup analysis this is likely due to the small sample size and the aim of this study was not to replicate these results.

There are several limitations of this study, including a small sample size, possible bias due to the retrospective nature of the study and the effects of longitudinal variability in scan acquisition on measurement of the semi-quantitative parameter, i.e. SUV_{max}. (given that scans were acquired over a long period of time). Further prospective studies are needed to address the impact of use of the Hopkins criteria on management.

Conclusion

The results of this study show that use of the Hopkins criteria for post-therapy assessment in patients with lung cancer represents an easy and reproducible method with substantial inter-observer agreement; this agreement approaches perfection for the classification of overall positive/negative residual disease status with no significant difference seen when determining the score with a semi-quantitative measure. Hopkins classification has a high PPV and accuracy and is easily understood by referring physicians.

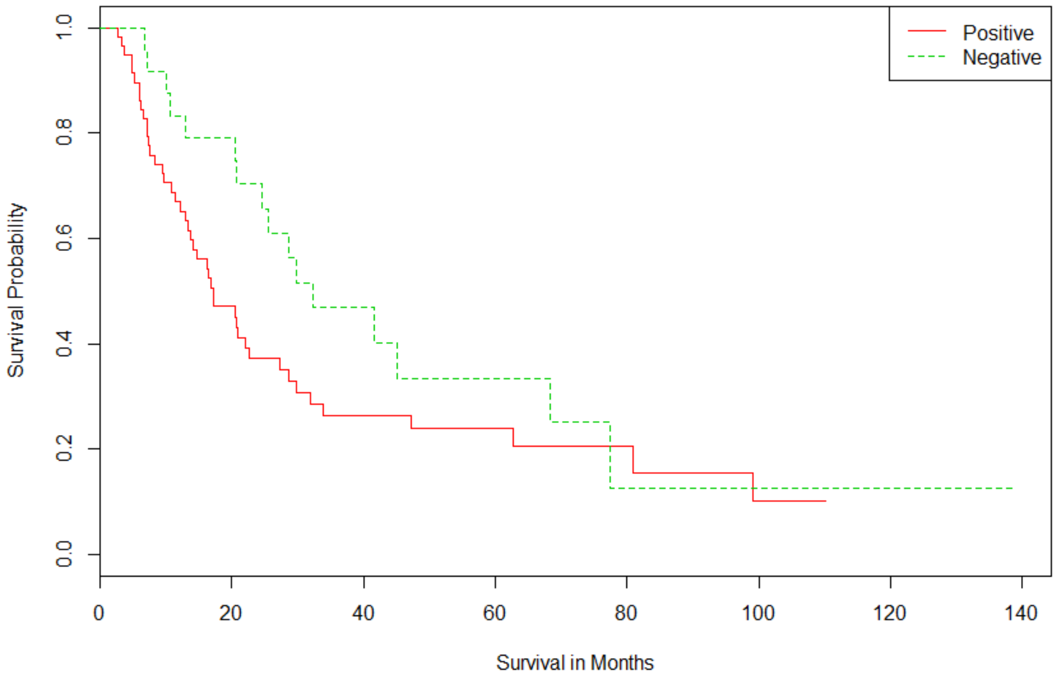


Figure 1. Kaplan-Meier survival plot for patients assessed as positive/negative for residual tumour using the Hopkins classification

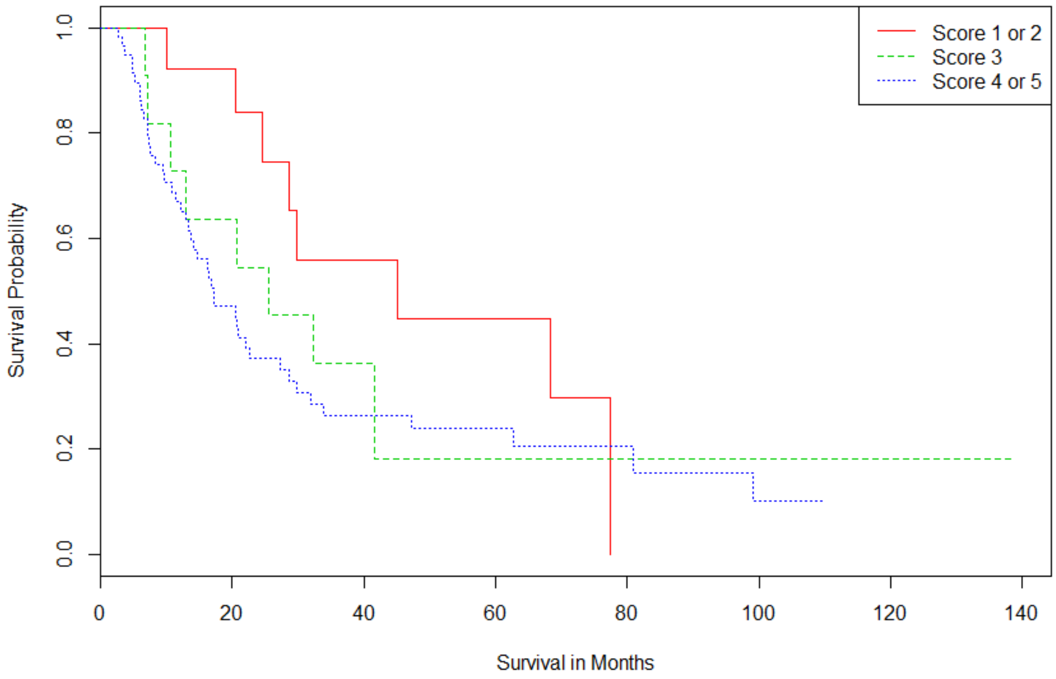


Figure 2. Kaplan-Meier survival plot for patients scored 1 or 2 vs 3 vs 4 or 5 on the Hopkins scoring system

Table 1. Hopkins Criteria qualitative post-therapy assessment scoring system

Score	Description	
1	Focal ^{18}F-FDG uptake visually less than or equal to mediastinal blood pool activity consistent with a complete metabolic response.	**Negative**
2	Focal ^{18}F-FDG uptake greater than mediastinal blood pool activity but less than liver representing a likely complete metabolic response.	
3	Diffuse ^{18}F-FDG uptake greater than mediastinal blood pool activity or liver uptake, representing likely inflammatory changes.	
4	Focal ^{18}F-FDG uptake greater than liver uptake, representing likely residual tumour.	**Positive**
5	Focal and intense ^{18}F-FDG uptake greater (2–3 times) than liver uptake was scored 5, consistent with residual tumour.	

Table 2. Patient characteristics

Characteristic	No.	%
Age (yr)		
≤40	2	2.4
41–60	32	37.6
>60	51	60
Sex		
Female	40	47.1
Male	45	52.9
Histology		
SCLC	10	11.8
NSCLC	74	87.1
Unspecified	1	1.2
History of smoking (+)	55	64.7
Stage		
I	10	11.8
II	9	10.6
III	32	37.6
IV	34	40.0
Surgery	15	17.6
Chemotherapy or immunotherapy	41	41.2
Radiotherapy	12	14.1
Surgery and chemoradiation	4	4.7
Chemoradiation	13	15.3
Interval between treatment and PET study (wk)		
0–8	51	60.0
8–12	13	15.3
12–24	21	24.7
PET/CT results		
Negative	26	69.4
Positive	59	30.6
Outcome (death)	60	70.6

Table 3. Diagnostic accuracy of the Hopkins scoring system

PET/CT results	**Disease negative**[a]	**Disease positive**[a]	**Total**
Negative	19	7	26
Positive	5	54	59
Total	**24**	**61**	**85**

[a] As assessed by imaging (n=66), histology (n=10) or clinical follow-up (n=9)

Acknowledgments: This study was supported by the NIHR Biomedical Research Centre University College London Hospital.

Compliance with Ethical Standards:

Disclaimer None

Funding None

Conflict of interest: The authors declare that they have no conflict of interest.

Ethical approval: Due to the retrospective character of the study, ethical approval was waived by the institutional ethics committee. The study conforms to the principles outlined in the Helsinki Declaration II.

Informed consent: All patients gave written informed consent with regard to the performed procedures and the fact that all data may be used for retrospective scientific analyses.

References

1. F. Bray, Jacques Ferlay, Isabelle Soerjomataram, Siegel; RL, Torre; LA, Jemal A (2018) Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 68:394–424

2. American Cancer Society (2019) Facts & Figures 2019. Am. Cancer Soc.

3. Uk CR (2014) Worldwide cancer statistics. In: Cancer Res. UK.

4. Alberts WM (2007) Follow up and surveillance of the patient with lung cancer: What do you do after surgery? Respirology 12:16-21

5. Sugimura H, Nichols FC, Yang P, Allen MS, Cassivi SD, Deschamps C, Williams BA, Pairolero PC (2007) Survival After Recurrent Nonsmall-Cell Lung Cancer After Complete Pulmonary Resection. Ann Thorac Surg 83:409–17

6. Sheikhbahaei S, Mena E, Marcus C, Wray R, Taghipour M, Subramaniam RM (2016) 18F-FDG PET/CT: Therapy Response Assessment Interpretation (Hopkins Criteria) and Survival Outcomes in Lung Cancer Patients. J Nucl Med 57:855–60

7. Van Tinteren H, Hoekstra OS, Smit EF, et al (2002) Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: The PLUS multicentre randomised trial. Lancet 359:1388–93

8. Fischer B, Lassen U, Mortensen J, et al (2009) Preoperative Staging of Lung Cancer with Combined PET–CT. N Engl J Med 361:32–9

9. Basu S, Kumar R, Ranade R (2015) Assessment of treatment response using PET. PET Clin 10:9-26

10. Wahl RL, Jacene H, Kasamon Y, Lodge MA (2009) From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors. J Nucl Med 50:122S-150S

11. O JH, Lodge MA, Wahl RL (2016) Practical PERCIST: A Simplified Guide to PET Response Criteria in Solid Tumors 1.0. Radiology 280:576–84.

12. Barrington SF, Kluge R (2017) FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas. Eur J Nucl Med Mol Imaging 44:97-110

13. Gallamini A, Barrington SF, Biggi A, et al (2014) The predictive role of interim positron

emission tomography for Hodgkin lymphoma treatment outcome is confirmed using the interpretation criteria of the Deauville five-point scale. Haematologica 99:1107–13

14. Mac Manus MP, Hicks RJ, Matthews JP, McKenzie A, Rischin D, Salminen EK, Ball DL (2003) Positron emission tomography is superior to computed tomography scanning for response-assessment after radical radiotherapy or chemoradiotherapy in patients with non-small-cell lung cancer. J Clin Oncol 21:1285–92

15. Mac Manus MP, Hicks RJ, Matthews JP, Wirth A, Rischin D, Ball DL (2005) Metabolic (FDG-PET) response after radical radiotherapy/chemoradiotherapy for non-small cell lung cancer correlates with patterns of failure. Lung Cancer 49:95–108

16. Vansteenkiste JF, Stroobants SG, De Leyn PR, Dupont PJ, Verbeken EK (1998) Potential use of FDG-PET scan after induction chemotherapy in surgically staged IIIa-N2non-small-cell lung cancer: A prospective pilot study. Ann Oncol 9(:1193–8

17. Soussan M, Chouahnia K, Maisonobe JA, Boubaya M, Eder V, Morère JF, Buvat I (2013) Prognostic implications of volume-based measurements on FDG PET/CT in stage III non-small-cell lung cancer after induction chemotherapy. Eur J Nucl Med Mol Imaging 40:668–76

18. Weber WA, Petersen V, Schmidt B, Tyndale-Hines L, Link T, Peschel C, Schwaiger M (2003) Positron emission tomography in non-small-cell lung cancer: Prediction of response to chemotherapy by quantitative assessment of glucose use. J Clin Oncol 21:2651–7

19. N. Plaxton, V. Moncayo, B. Barron RH (2014) Factors that influence standard uptake values in FDG PET/CT. J Nucl Med 55:1356–1356

20. Azmi NHM, Suppiah S, Liong CW, Noor NM, Said SM, Hanafi MH, Kaewput C, Saad FFA, Vinjamuri S (2018) Reliability of standardized uptake value normalized to lean body mass using the liver as a reference organ, in contrast-enhanced 18F-FDG PET/CT imaging. Radiat Phys Chem 147:35–39

Nuclear Medicine Communications

Author Ethics Checklist

*Please complete this form in Word by entering information where indicated by arrowheads***.** Full and informative answers are needed. This document is for the editors' quick reference; **please make sure you have also included relevant information in your submitted article.**

Ensure document is unprotected in "Tools" in the menu bar in order to write in it.

Title of paper:	► Revalidation of PET/CT criteria (Hopkins criteria) for the assessment of therapeutic response in lung cancer patients: Inter-reader reliability, accuracy and survival outcomes	
Names of authors:	► Khulood Al Riyami, Noor Al Nuaimi, Ruta Kliokyte, Andrew Thornton, Stefan Voo, Jamshed Bomanji and Francesco Fraioli	
Office use only	*Sub. no.:*	*Vol/Issue:*

1. DUPLICATE PUBLICATION includes papers, or letters to the Editor previously published in this, or another journal. Abstracts of papers presented at meetings and published in the proceedings of such meetings do not constitute duplicate publication, but should be disclosed by including a note at the beginning of the paper, i.e. "Data presented previously at (state meeting) and published as abstract in (give reference)". Have you published these data previously? If so, Have you acknowledged this?

► This is an original paper and has not submitted or presented elsewhere.

2. CONFLICT OF INTEREST includes financial support from the biomedical industry or other commercial sources in the form of research grants, bench fees, consultancy or lecture fees, travelling expenses, payment of registration fees, consultancy appointments, posts held in the biomedical industry or equipment manufacturers, stock holdings in the company, free supply of drugs and the like. These should be assessed in relation to each author.
If conflict of interest is present, we will publish a statement to that effect at the end of your paper, unless you have valid objections to this. Have any of the authors any conflict of interest? Please state details.

► No conflict of interest fo all authors.

3. CONSENT Please confirm that:

- all appropriate subjects' consents have been obtained prior to submission of your manuscript.
- any information that identifies a patient has been removed from any images you have submitted.

► all patient's information has been removed. This is a retrospective, single-centre cohort diagnostic study with patient inclusion conforming to the principles outlined in the Helsinki Declaration II. Due to the retrospective character of the study, ethical approval was waived by the institutional ethics committee.

4. ETHICS All research studies need to be approved by the local Research Ethics Committee. Was your study? If you feel that Ethical committee approval is not required, please give reasons.

► See above.

5. AUTHORS CONTRIBUTIONS Please state briefly how each of the authors contributed to the study, to data analysis and to the writing of your paper. For a person to qualify as an author, their contribution should be sufficient for them to assume responsibility for the study.

► All authors contribuited equally to the manuscript, by analysing data, collecting clinical information, revising the results and drafting the manuscript.

6. STATISTICAL ANALYSIS Kindly please let me know who performed the statistical analysis of your data.

► Dr. Andrew Thornton, and Stefan Voo have a specific backround in medical statistic and perform the statistical analysis.	
Other information for the Editor that may be relevant: ►	
Name of person completing this form: *(Please print if handwritten)*	► Francesco Fraioli
Date:	► 31/7/2019

RESPONSE TO THE REVIEWERS (REBUTTAL LETTER)

>The authors would like to thank the reviewers for their constructive comments which helped us to improve the quality of our manuscript. The issues raised by the reviewers and editor were carefully addressed.

Reviewer #1:

1.A limitation of this study is in the heterogeneity of patients included. The authors included patients with stage I to IV lung ca who were treated by different modalities. This group represents patients with disease which has varying biology and treated with different intent, curative versus palliative. Unfortunately, the authors did not consider to evaluate for the impact of disease stage on survival as they did for tumor histology and treatment modality. The modest study population may not allow for this subgroup analysis. This point needs to be emphasize as a significant limitation of the study.

> Much thanks to the reviewer for this excellent observation. We have added this to the list of study limitations .(page 11, paragraphs 3, lines 11-15, highlighted in red).

2.Time of FDG PET/CT for response assessment varied among patients (imaging done over about 6 months since completion of treatment). Unfortunately, the author calculated OS survival from the time FDG PET/CT was obtained. This variability in the time between time of completion of therapy and FDG PET/CT imaging could have affected the survival data obtained in this study.

> The method by which OS was calculated was that used in the original study in order to compare the results with the current revalidation study. We have added this point in the limitation section. (page 11, paragraph 3, lines 19-21, highlighted in red).

3. Histological correlation of image finding was obtained in 11.8%. This is a major limitation and should be included in the paragraph discussing the limitations of this study.

> This has been added in the limitation paragraph. (page 11, paragraphs 3, lines 15-16, highlighted in red).

4.Sufficient details are not provided regarding how the semi-quantitative criteria was computed. Please provide details on this. Alternatively, the authors may reference a published study where such details have been presented.

> This has been further clarified in the methodology section. Please see (page 5, paragraph 6, lines 21-26, highlighted in red)

Francesco and Andy its not clear to me if the reviewer means to explain the methodology more , or he/she wants the physics details behind how SUVmax was calculated. If its related to physics can you please check with the specification of the machine.

5."There was, however, a significant difference in OS based on Hopkins criteria for positivity, with Score (logrank) test = 11.0 on 4 df, p=0.027" This reviewer does not understand what the authors meant by this statement especially since it was mentioned earlier that that was no significant difference in OS based on Hopkins positivity, please rephrase.

> This line has been further clarified and should have read: There was, however, a significant difference in OS based on the comparison of the individual Hopkins criteria score groupings with each other, with Score (logrank) test = 11.0 on 4 df, p=0.027

6."PPV and accuracy and is easily understood by referring physicians" Unfortunately no data supporting this statement was presented in this study. Please remove.

- ➢ This has been re-phrased to " we believe can be easily understood". Please see . (page 11 paragraphs 4, lines 29,).

7.Please include the respective logrank test results, the hazard ratio with the 95%CI and the p values within the survival curves.

- ➢ These have been added to the captions for the two images.

8.Please provide more details on the areas of discrepancies in interpretation between the two readers. This can be summarized using a table.

- ➢ Have added a table summarizing the different scorings between the two observers.

9.Include representative images showing positive and negative PET findings.

- ➢ Done

10. "Due to the retrospective character of the study, ethical approval was waived by the institutional ethics committee." I guess, here, the authors meant to say that their institutional ethics committee waived obtaining formal informed consent from the patients due to the retrospective nature of this study rather than "ethical approval was waived". Please rephrase.

- ➢ Thank you for highlighting this point, this has been rephrased

----

Reviewer #2:

1. In addition to qualitative Hopkins criteria, you measured the five-point scoring system using SUVmax as a semi-quantitative measure of tracer uptake. How did you do it? What was the SUV values? You should place the SUV values in the text. You should generate a table for both data of qualitative Hopkins criteria and data of semi-quantitative Hopkins criteria.

> This has been further clarified in the methodology section. Please see (page 5, paragraph 6, lines 21-26, highlighted in red).

With regards to the SUVmax values , the authors felt that absolute values were beyond this study's aim, but rather how they compare to the background values (liver and mediastinal pool) in order to categorise them into the 5 point scale .. However we plan to publish a further study looking at the SUVmax values and its impact on the overall survival in view of the Hopkins crietria.

The data for both qualitative Hopkins criteria and data of semi-quantitative Hopkins criteria are within the text in the respective results section (pages 7 & 8)

2. Was there was a significant difference in OS between patients with positive and negative Hopkins score both in those who had surgical resection as part of the primary treatment and in those who were treated with chemotherapy with or without radiation?

- This has been described in the results section under "**Survival Outcomes: Impact of Tumour Histology and Treatment Modality",** in which there was a difference in the OS between the positive and negative Hopkins score sub analysed according to treatment, with the difference mainly noted in those treated with radiotherapy or surgery, however this was not statistically significant. However when subanalysing according to the 5 point score there was this difference was statistically significant in those with score 5 verses those who scored 4 or less when treated with radiotherapy or surgery. Please see . (page 10, paragraphs 1).

Title: Revalidation of PET/CT criteria (Hopkins criteria) for the assessment of therapeutic response in lung cancer patients: Inter-reader reliability, accuracy and survival outcomes

Short title: Revalidation of Hopkins criteria in lung cancer

Authors: Khulood Al Riyami[1,2], Noor Al Nuaimi[1], Ruta Kliokyte[3], Stefan Voo[1], Andrew Thornton[1], Jamshed Bomanji[1] and Francesco Fraioli[1]

Affiliation: 1. Institute of nuclear medicine, University College Hospital London, UK

2.Department of radiology and molecular imaging, Sultan Qaboos University Hospital, Oman

3.Centre of Radiology and Nuclear Medicine, Vilnius University Santaros clinics, Lithuania

Correspondence to: Francesco Fraioli

Number or words: 4210

Number of tables: 3

Number of figures: 2

Abstract

Background/Aim. Systematic reporting using qualitative evaluation of PET/CT results has been demonstrated to be very accurate and reproducible in post-therapy assessment of lung cancer (so-called Hopkins criteria). Our aim was to test, in a different cohort of patients, the Hopkins criteria for assessment of therapeutic response in lung cancer and to compare the results with those obtained using a semi-quantitative evaluation of uptake.

Methods. This is a retrospective study. A total of 85 patients with known lung cancer who underwent ^{18}F-FDG PET/CT assessment within 24 weeks (mean 7.9 weeks) of completion of treatment were included. Treatments included surgical resection, chemotherapy, radiation therapy, immunotherapy or combinations thereof. PET/CT interpretation was done by two nuclear medicine physicians, and discrepancies were resolved by a third interpreter. Studies were scored both according to the Hopkins criteria using qualitative assessment of tracer uptake for the primary tumour, loco-regional disease in the mediastinum and distant metastatic sites and by applying the same 5-point score using a semi-quantitative measure, SUV$_{max}$. Overall scores of 1, 2 and 3 were considered negative for residual disease, while scores of 4 and 5 were considered positive. Patients were followed up for a median of 18.5 months (range 2–139 months). Kaplan-Meier plots with a Mantel-Cox log-rank test were performed, considering death as the endpoint. Inter-reader variability was assessed using percent agreement and kappa statistics.

Results. The Cohen κ coefficient analysis showed substantial agreement between the two interpreters on the five-point Hopkins criteria scoring, with a κ of 0.73. There was almost perfect agreement between the interpreters with respect to classification as positive or negative according to the Hopkins criteria, with a κ of 0.89. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the Hopkins criteria were 88.5% (95%CI 80.6%–96.5%), 79.2% (95%CI 63.2%–95.1%), 91.5% (95%CI 84.4%–98.6%), 73.1% (95%CI 61.8%–84.4%) and 85.9% (95%CI 78.5%–93.3%) respectively. There was almost perfect agreement between the qualitative and semi-quantitative scoring with a κ of 0.87, with sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the semi-quantitative Hopkin's criteria of 86.9% (95% CI 78.4%–95.4%), 79.2% (95%CI 62.9%–95.4%), 91.4% (95%CI 84.2%–98.6%), 70.4% (95%CI 58.6%–82.1%), and 84.7% (95%CI 80.8%–92.4%) respectively.

Conclusion. The use of Hopkins criteria for post-therapy assessment in patients with lung cancer represents an easy and reproducible method with substantial to almost perfect inter-observer agreement and high positive predictive value and accuracy; moreover, it is easily understood by referring physicians. Additionally, there was no significant difference when applying a semi-quantitative measure to the same 5 point score.

Key Words. PET/CT; lung cancer; therapy assessment; Hopkins criteria

Introduction

Lung cancer is the leading cause of cancer death among men and the second leading cause of cancer death among women worldwide, with over 2 million new cases in 2018 [1]. In the United states, the estimates for lung cancer for 2019 is over 200,000 new cases and over 140,000 deaths from lung cancer [2]. Similarly, in the United Kingdom, it was the most common cause of cancer death in 2014 [3]. Despite the advances in the available therapeutic options, recurrences after lung cancer surgery typically happen rapidly, with 90%–95% occurring within 5 years [4]. Survival time following local or distant recurrence averages less than a year, including among patients who receive salvage treatment [5]. In this context, the assessment of therapeutic response could change the clinical management of the patient and potentially improve survival [6].

Fluorine-18 fluorodeoxyglucose (FDG) PET/CT has been established as an important imaging method for the diagnostic work-up of lung cancer patients [7, 8], though the need for an established systematic and reproducible interpretation system has also been recognised [9]. Evaluation of response using PET can be performed visually or semi quantitatively through comparison of a lesion's SUV max with the background liver uptake and the mediastinal blood pool.

The recently introduced Hopkins criteria for lung cancer response assessment in PET/CT seems to offer substantial inter-observer agreement as well as high sensitivity and specificity to predict survival in lung cancer patients, irrespective of tumour histology and treatment [6]. The Hopkins criteria does not suggest comparison using the SUV_{max}; although there are other standards such as PERCIST, those are of not easily adopted in clinical practice, hence we will consider the semi-quantitative extension to the Hopkin's criteria where assessment of uptake will be compared against the SUV_{max}.

The aim of this study was to externally validate the reproducibility of the Hopkins criteria and compare their use with a semi-quantitative method of evaluation of uptake.

Materials and Methods

This is a retrospective, single-centre cohort diagnostic study with patient inclusion conforming to the principles outlined in the Helsinki Declaration II. Due to the retrospective character of the study, ethical approval was waived by the institutional ethics committee.

Eligible Patients and Follow-up

A retrospective evaluation was conducted on consecutive patients with lung cancer who had undergone PET/CT scans for post-therapy assessment.

Inclusion criteria. The study included patients with biopsy-proven lung cancer who underwent post-therapy ^{18}F-FDG PET/CT at our institute between 2007 and 2015 for the evaluation of treatment response. All patients were treated with surgical resection, chemotherapy, radiotherapy or immunotherapy, or a combination of these treatments, and then underwent ^{18}F-FDG PET/CT within 24 weeks of therapy.

Patient demographics, clinical history and clinical data were collected from the electronic medical records for up to 6 months following PET/CT.

Exclusion criteria. Patients with no available follow-up, patients with PET/CT imaging >24 weeks post therapy, patients with no available imaging and those with concurrent malignancies were excluded.

Image Analysis

Hopkins Criteria for Post-Therapy Assessment on PET/CT.

Studies were scored using a qualitative five-point scale, for the primary tumour, locoregional disease in the mediastinum and distant nodal or metastatic sites (Table *1)*. The visual activity in the mediastinal blood pool and in the liver was taken as the background blood pool for reference.

Semi-quantitative Criteria for Post-therapy Assessment on PET/CT

The same five-point scale (Table 1) was applied using SUV$_{max}$ values as a semi-quantitative measure of tracer. SUVmax values were measured in the mediastinal blood pool (at the aortic arch,sparing the vessel walls), liver background (right lobe, excluding regions that were involved by disease) and the highest SUV$_{max}$ value within the sites of active disease whether it being in the primary tumour, lymph nodes or distant metastasis were recorded and categorised according to the five point scale.

Definition of Positive and Negative PET/CT Studies. On the basis of the qualitative five-point scale, the studies were grouped as positive or negative for the primary tumour, locoregional disease in the mediastinum and distant metastatic lesions. Overall assessment was denoted by the overall score, which was the highest score among the scores for the primary tumour and locoregional and distant metastatic lesions, if present. Scores 1, 2 and 3 were considered negative for residual tumour and scores 4 and 5 were considered positive.

All PET/CT studies were retrieved from the institutional Picture Archiving and Communication System (PACS) and reviewed on a Carestream Vue PACS workstation/viewing platform (version 12.1.5.7014, Carestream Health Inc). All images were interpreted by two reviewers independently (readers 1 and 2). Reader 1 was a board-certified radiologist with more than 4 years' subspecialty training in nuclear medicine, and reader 2 was a board-certified medical physician with subspecialty in nuclear medicine and expertise in lung oncologic imaging. In order to reach a consensus, any discrepancies were resolved by a third interpreter, a senior consultant dual board-certified in nuclear medicine and radiology who was the main nuclear medicine representative at the local lung cancer multidisciplinary meeting (MDM). This final consensus score was used to determine the final Hopkins score.

Outcome measures

As in the original article on use of the Hopkins criteria (6), histological confirmation of PET/CT-positive lesions, alternative imaging modalities or clinical follow-up of 6 months after PET/CT were considered as the reference standard.

The sensitivity and specificity, positive predictive value, negative predictive value and accuracy of the post-therapy PET/CT assessment criteria, along with 95% confidence intervals (CIs), were calculated by constructing a 2×2 contingency table (cross-relating PET/CT results of the reference standards).

Overall survival for all cases was defined as the time interval in months between the date of the post-therapy PET/CT study and the date of death or loss to follow-up. The date of the study was recorded from the radiology information system (RIS) and the date of death was extracted from the electronic medical records.

Statistical analysis

Descriptive values are presented as mean (with standard deviation) or median (with 25th to 75th percentile range) if the data were not normally distributed. Categorical variables are presented as frequency (with percentage). The Cohen κ co-efficient was calculated to measure inter-interpreter agreement and inter-criteria agreement. Survival probabilities were generated using Kaplan-Meier survival curves and compared using the Mantel-Cox log-rank test. Univariate and multivariate Cox regression analyses were performed, considering death as the endpoint. Subgroup analysis was performed to assess the impact of histological subtype and prior treatment on the prognostic value of the Hopkins score. The statistical significance level was set at a p value of less than 0.05. Statistical analysis was performed using R 3.3.1.

Results

Patient Characteristics and Follow-up

Eighty-five patients were included in the study (45 male, 40 female; mean age ± SD, 63.5±10.2 years). A history of smoking was present in 55 patients (64.7%). The histological subtype of the primary malignancy was identified as small cell lung cancer (SCLC) in 10 patients (11.8%) and non-small cell lung cancer (NSCLC) in 74 (87.1%); histological subtype was not identified in one patient. The demographic details of the 85 patients included in the study are summarised in Table *2*. The median follow-up of these patients was 18.5 months (range 2–139 months) after completion of the post-treatment assessment PET/CT.

Time Interval to Post-treatment PET/CT

Post-treatment ^{18}F-FDG PET/CT for assessment of therapeutic response was performed between 0 and 24 weeks after completion of treatment. The average interval between the date of completion of treatment and the post-treatment ^{18}F-FDG PET/CT was 7.9 weeks (median 6 weeks, range 0–24 weeks).

Interpreter Classification of PET/CT studies (by Hopkins)

On the basis of the final Hopkins qualitative criteria scores, PET/CT studies were characterised as positive in 59 patients (69.4%) and as negative in 26 (30.6%). On the 59 positive PET/CT studies, the most avid residual disease was identified at the primary site in 27 patients (45.8%), at sites of nodal disease in 23 (39.0%), at sites of distant metastases in 7 (11.9%) and in the pleura in 2 (3.4%). Of the 26 PET/CT studies characterised as negative, 14 (53.6%) were scored as 1 or 2 and 12 (46.2%) as 3.

The accuracy of the scoring system was assessed by imaging in 66 cases (77.6%), histology in ten (11.8%) and clinical follow-up in nine (10.6%). Table 3 summarises the results of this follow-up. The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the scoring system were 88.5% (95%CI 80.6%–96.5%), 79.2% (95%CI 63.2%–95.1%), 91.5% (95%CI 84.4%–98.6%), 73.1% (95%CI 61.8%–84.4%) and 85.9% (95%CI 78.5%–93.3%) respectively.

Interpreter Classification of PET/CT studies (by semi-quantitative Hopkins extension)

On the basis of the final semi-quantitative criteria scores, PET/CT studies were characterised as positive in 58 patients (68.2%) and as negative in 26 patients (30.6%). One patient had to be excluded from assessment as height and weight were not provided and SUV_{max} could not therefore be calculated. On the 58 positive PET/CT studies, the most avid residual disease was identified at the primary site in 27 patients (46.5%), at sites of nodal disease in 22 (37.9%), at sites of distant metastases in 7 (12.1%) and in the pleura in 2 (3.4%). Of the 26 PET/CT studies characterised as negative, 14 (53.6%) were scored as 1 or 2 and 12 (46.2%) as 3.

The accuracy of the scoring system was assessed by imaging in 66 cases (78.6%), histology in nine (10.7%) and clinical follow-up in nine (10.7%). The sensitivity, specificity, positive predictive value, negative predictive value and accuracy of the semi-quantitative Hopkin's criteria were 86.9% (95% CI 78.4%–95.4%), 79.2% (95%CI 62.9%–95.4%), 91.4% (95%CI 84.2%–98.6%), 70.4% (95%CI 58.6%–82.1%), and 84.7% (95%CI 80.8%–92.4%) respectively.

Interobserver agreement

The Cohen κ coefficient analysis indicated substantial agreement between the two interpreters (R1 and R2) on the five-point qualitative Hopkins criteria scoring, with a κ of 0.73. (Table *4*). Discrepancies between the two interpreters (15 patients, 17.2%) were resolved by the third interpreter (R3). There was almost perfect agreement between R1 and R2 in terms of positive versus negative classification according to the Hopkins criteria, with a κ of 0.89 (discrepancies occurred in only four patients, 4.7%). When scoring was performed using the semi-quantitative measure, SUV_{max}, substantial agreement was again observed between the two interpreters, with a κ of 0.72, but discrepancies occurred in 19 patients (22%), i.e. four more than in the qualitative assessment. (Table *5*).

Intercriteria agreement

There was also almost perfect agreement between the final five point qualitative and semi-quantitative scoring with a κ of 0.87 – this increased to κ of 0.97 when scoring was qualified as positive versus negative with only 1 discrepancy (1.2%). 1 patient had to be excluded from the quantitative analysis as SUVs could not be generated due to a lack of patient height and weight information.

Survival Outcome in All Patients

The median follow-up of the study population was 20.5 months (range 2–139 months) after completion of the post-treatment assessment PET/CT and 60 patients died during follow-up (70.6%). The median survival of the Hopkins criteria-positive group was 17.2 months (95% CI 13.7–28.6 months) and 44 (74.6%) patients died in this group. The median survival of the Hopkins criteria-negative group was 32.4 months (95% CI, 24.6–), and 16 (61.5%) patients died in this group. The Kaplan-Meier survival analysis showed a non-significant difference in overall survival (OS) (Score (logrank) test = 2.66 on 1 df, p=0.1027), with a hazard ratio (HR) of 1.61 (95% CI 0.90–2.86) (Figure *1*). Hopkins criteria positivity was not statistically significant in multivariate analysis when compared with stage, age or sex. Results were similar for the semi-quantitative scoring system.

In overall assessment using the five-point interpretation scale, there was no significant in difference in OS between patients who scored 1 or 2 (n=15) versus those who scored 3 (n=13) versus those who scored 4 or 5 (n=59), (Score (logrank) test = 3.06 on 2 degrees of freedom [df], p=0.2162). There was, however, a significant difference in OS based on the comparison of the individual Hopkins criteria score groupings with each other, with Score (logrank) test = 11.0 on 4 df, p=0.027 and there was a significant difference in OS if those with Score 5 (n=44) were compared against those scoring less than 5 (n=41), Score (logrank) test = 10.43 on 1 df, p=0.001243 HR 2.4 (95%CI 1.38–4.03). Results were similar for the semi-quantitative scoring system.

Survival Outcomes: Impact of Tumour Histology and Treatment Modality

Based on tumour histology, ten patients (11.8%) were diagnosed with SCLC and 74 (87.1%) with NSCLC (one case was unspecified). The Kaplan-Meier analysis showed a difference between OS in the eight SCLC patients who had a positive PET/CT result (median survival 6.75 months) and in the two who had a negative PET/CT result (median survival 10.61 months), with an HR of 2.83; however, this difference was not statistically significant due to the small numbers of patients involved. In patients with NSCLC (n=74) or an unspecified histology (n=1), OS was shorter in those with a positive PET/CT (n=51, median survival 20.6 months) than in those with a negative PET/CT (n=24, median survival 32.4 months), with an HR of 1.44; again, however, this result was not statistically significant. When restricted to those patients who scored 5, the median survival difference for SCLC was 5.11 months versus 10.9 HR 2.25, again non-significant, but was significant for NSCLC (or unspecified) with 14.7 months versus 33.8, HR 2.18 (95%CI 1.23–3.89) Score (logrank) test = 7.36 on 1 df, p=0.00665. Results were similar for the semi-quantitative scoring system.

There was no significant difference in OS when patients were distinguished according to preceding treatment; however, there was again a trend towards a difference in survival between those with a positive and those with a negative PET/CT by Hopkins Criteria, the difference being greater in those who were treated with radiotherapy or surgery [median OS 16.8 months vs 25.4 months (positive vs negative PET/CT, respectively) for those treated with chemotherapy or immunotherapy alone, compared with 18.9 months vs 45.1 months for those who received radiotherapy or surgery]. Again when restricted to those patients who scored 5 there is a non-significant but almost significant difference in survival between those where treated with chemotherapy or immunotherapy alone (median OS 11.4 months vs 33.8 months), but there is a significant difference in survival with those with score 5 treated with radiotherapy and surgery versus those with scores of 4 or less, (median OS 14.2 vs 32.4 months, HR 2.4 (95%CI 1.14–5.08) Score (logrank) test = 5.69 on 1 df, p=0.01705. Results were similar for the semi-quantitative scoring system although the difference between survival at score 5 vs <5 is significant in those treated with chemotherapy or immunotherapy alone.

Discussion

The role of ^{18}F-FDG PET/CT in treatment response assessment has been widely established in other types of malignancy such as lymphoma, using the Deauville criteria [12, 13].

The use of ^{18}F-FDG PET/CT in lung cancer staging is well established and demonstrated that PET response assessment is much more strongly correlated with survival than response measured by CT scanning [14]; however there is less well established consensus or recommendations on which reporting system is best to use, either visual or quantitative. In our study, we sought to answer two questions, whether the results from visual analysis were reproducible, as previously demonstrated, and second if there is an agreement between the visual and quantitative scoring systems.

In our study we can confirm the results shown by Sheikhbahaei et al. that the Hopkins criteria permit reproducible qualitative assessment of therapeutic response using visual ^{18}F-FDG uptake and can be of a great value for patient care. When using the qualitative five-point Hopkins scoring system, similar to the data presented [6] we observed substantial agreement between the readers and almost perfect agreement when categorizing the patients into positive and negative for disease as per the criteria.

Several studies have demonstrated the added value of post-treatment ^{18}F-FDG PET in the prognostication of patients with lung cancer. These studies reported longer survival in patients with a complete metabolic response, post-therapy reduction in ^{18}F-FDG uptake and changes in total lesion glycolysis and metabolic tumour volume [7, 14–18]. As observed by

Sheikhbahaei et al [6], there was a trend towards better overall survival in patients with negative post-treatment scans and in those with scores of 1 and 2; however, in our study these differences were not statistically significant, which may be attributed to the population size.

As for the second purpose, in our study we demonstrated no significant difference in the inter-reader and inter criteria agreement using the qualitative Hopkins criteria and the same five-point scoring system using SUV$_{max}$ as a semi-quantitative measure of tracer uptake. This highlights the fact that the simplified method of using visual assessment for scoring is a reliable technique.

Furthermore, to strengthen the visual assessment criteria it should be recognised that SUV$_{max}$ values can be affected by patient related factors such as fasting blood glucose levels, altered bio-distribution of 18F-FDG which can occur in morbid obesity and technical parameters such as varying uptake time, image noise, partial volume effect and differences in acquisition techniques such as the number of iterations [19, 20].

Finally, although we did not find significant difference in OS when restricted to subgroup analysis this is likely due to the small sample size and the aim of this study was not to replicate these results.

There are several limitations of this study, the most significant being the heterogeneity of the patients which included different disease stages, histology and treatment modalities, however subgroup analysis would have been limited due to the modest study population. Additional limitations included a small sample size, availability of histological correlation in a limited number of patients, possible bias due to the retrospective nature of the study and the effects of longitudinal variability in scan acquisition on measurement of the semi-quantitative parameter, i.e. SUV$_{max}$. (given that scans were acquired over a long period of time). Furthermore as in the original study, the OS was calculated from the time FDG PET/CT was obtained, however the time of FDG PET/CT for response assessment varied among the patients and this could have affected the survival data obtained. Further prospective studies are needed to address the impact of use of the Hopkins criteria on management.

Conclusion

The results of this study show that use of the Hopkins criteria for post-therapy assessment in patients with lung cancer represents an easy and reproducible method with substantial inter-observer agreement; this agreement approaches perfection for the classification of overall positive/negative residual disease status with no significant difference seen when determining the score with a semi-quantitative measure. Hopkins classification has a high PPV and accuracy and we believe can be easily understood by referring physicians.

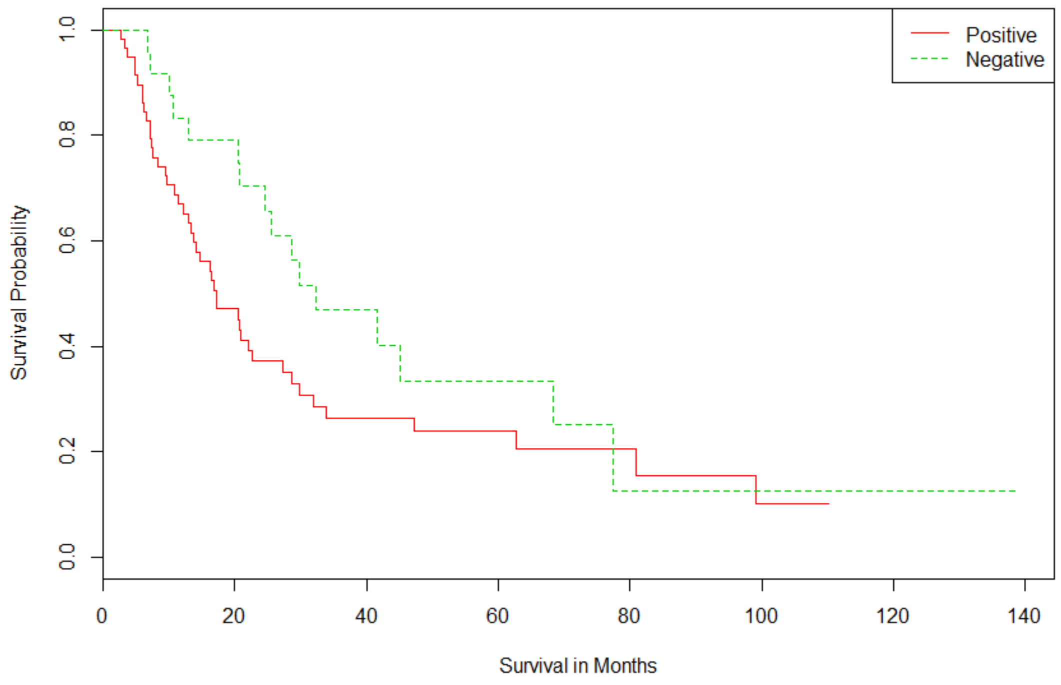


Figure 1. Kaplan-Meier survival plot for patients assessed as positive/negative for residual tumour using the Hopkins classification showed a non-significant difference in overall survival (OS) (Score (logrank) test = 2.66 on 1 df, p=0.1027), with a hazard ratio (HR) of 1.61 (95% CI 0.90–2.86)

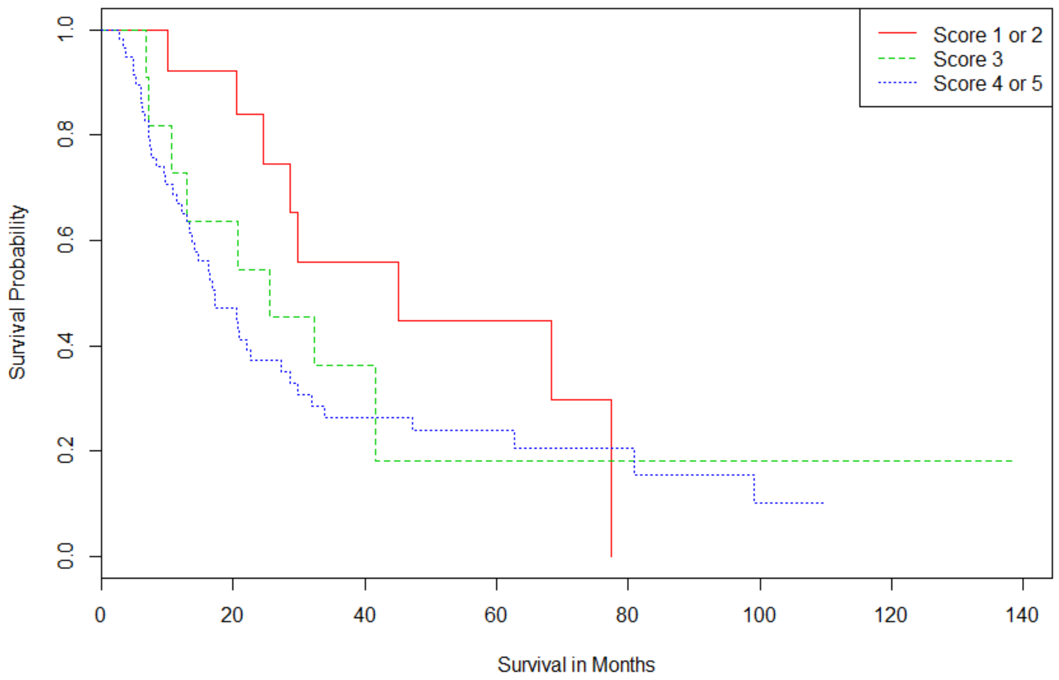


Figure 2. Kaplan-Meier survival plot for patients scored 1 or 2 vs 3 vs 4 or 5 on the Hopkins scoring system also showed a non-significant difference in overall survival (OS) (Score (logrank) test = 3.06 on 2 degrees of freedom [df], p=0.2162)

Table 1. Hopkins Criteria qualitative post-therapy assessment scoring system

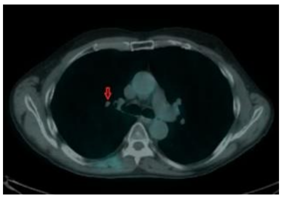
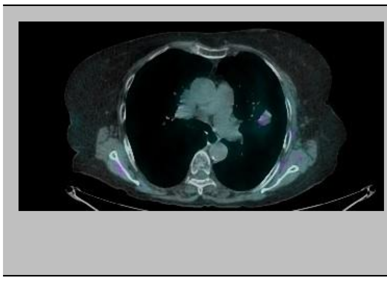
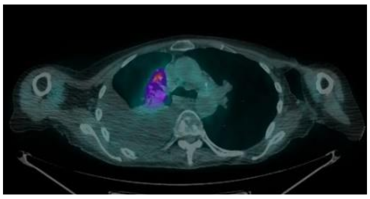
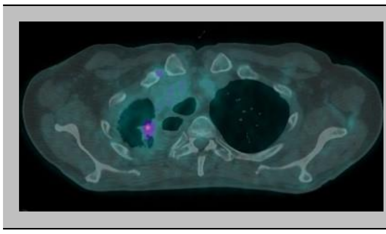
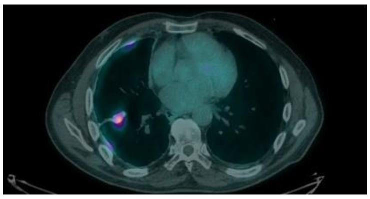
	Score	Description	
	**1**	Focal ^{18}F-FDG uptake visually less than or equal to mediastinal blood pool activity consistent with a complete metabolic response.	**Negative**
	**2**	Focal ^{18}F-FDG uptake greater than mediastinal blood pool activity but less than liver representing a likely complete metabolic response.	
	**3**	Diffuse ^{18}F-FDG uptake greater than mediastinal blood pool activity or liver uptake, representing likely inflammatory changes.	
	**4**	Focal ^{18}F-FDG uptake greater than liver uptake, representing likely residual tumour.	**Positive**
	**5**	Focal and intense ^{18}F-FDG uptake greater (2–3 times) than liver uptake was scored 5, consistent with residual tumour.	

Table 2. Patient characteristics

Characteristic	No.	%
Age (yr)		
≤40	2	2.4
41–60	32	37.6
>60	51	60
Sex		
Female	40	47.1
Male	45	52.9
Histology		
SCLC	10	11.8
NSCLC	74	87.1
Unspecified	1	1.2
History of smoking (+)	55	64.7
Stage		
I	10	11.8
II	9	10.6
III	32	37.6
IV	34	40.0
Surgery	15	17.6
Chemotherapy or immunotherapy	41	41.2
Radiotherapy	12	14.1
Surgery and chemoradiation	4	4.7
Chemoradiation	13	15.3
Interval between treatment and PET study (wk)		
0–8	51	60.0
8–12	13	15.3
12–24	21	24.7
PET/CT results		
Negative	26	69.4
Positive	59	30.6
Outcome (death)	60	70.6

Table 3. Diagnostic accuracy of the Hopkins scoring system

PET/CT results	**Disease negative**[a]	**Disease positive**[a]	**Total**
Negative	19	7	26
Positive	5	54	59
Total	**24**	**61**	**85**

[a] As assessed by imaging (n=66), histology (n=10) or clinical follow-up (n=9)

Table 4. Interobserver Agreement for the Hopkins Criteria

	R1	**Negative**			**Positive**	
R2		**1**	**2**	**3**	**4**	**5**
Negative	**1**	**7**	0	3	0	0
	2	0	**2**	1	2	0
	3	1	2	**7**	1	0
Positive	**4**	0	0	0	**12**	2
	5	0	0	1	2	**42**

Table 5. Interobserver Agreement for the Semi-quantitative Hopkins Criteria

	R1	**Negative**			**Positive**	
R2		**1**	**2**	**3**	**4**	**5**
Negative	**1**	**7**	0	3	0	0
	2	0	**3**	2	0	0
	3	0	2	**7**	1	0
Positive	**4**	0	0	0	**13**	4
	5	0	0	1	1	**40**

Acknowledgments: This study was supported by the NIHR Biomedical Research Centre University College London Hospital.

Compliance with Ethical Standards:

Disclaimer None

Funding None

Conflict of interest : The authors declare that they have no conflict of interest.

Ethical approval and consent: All patients gave written informed consent with regard to the performed procedures and the fact that all data may be used for retrospective scientific analyses. As this is a retrospective study, further explicit ethical approval was waived by the institutional ethics committee. The study conforms to the principles outlined in the Helsinki Declaration II.

References

1. F. Bray, Jacques Ferlay, Isabelle Soerjomataram, Siegel; RL, Torre; LA, Jemal A (2018) Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 68:394–424

2. American Cancer Society (2019) Facts & Figures 2019. Am. Cancer Soc.

3. Uk CR (2014) Worldwide cancer statistics. In: Cancer Res. UK.

4. Alberts WM (2007) Follow up and surveillance of the patient with lung cancer: What do you do after surgery? Respirology 12:16-21

5. Sugimura H, Nichols FC, Yang P, Allen MS, Cassivi SD, Deschamps C, Williams BA, Pairolero PC (2007) Survival After Recurrent Nonsmall-Cell Lung Cancer After Complete Pulmonary Resection. Ann Thorac Surg 83:409–17

6. Sheikhbahaei S, Mena E, Marcus C, Wray R, Taghipour M, Subramaniam RM (2016) 18F-FDG PET/CT: Therapy Response Assessment Interpretation (Hopkins Criteria) and Survival Outcomes in Lung Cancer Patients. J Nucl Med 57:855–60

7. Van Tinteren H, Hoekstra OS, Smit EF, et al (2002) Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: The PLUS multicentre randomised trial. Lancet 359:1388–93

8. Fischer B, Lassen U, Mortensen J, et al (2009) Preoperative Staging of Lung Cancer with Combined PET–CT. N Engl J Med 361:32–9

9. Basu S, Kumar R, Ranade R (2015) Assessment of treatment response using PET. PET Clin 10:9-26

10. Wahl RL, Jacene H, Kasamon Y, Lodge MA (2009) From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors. J Nucl Med 50:122S-150S

11. O JH, Lodge MA, Wahl RL (2016) Practical PERCIST: A Simplified Guide to PET Response Criteria in Solid Tumors 1.0. Radiology 280:576–84.

12. Barrington SF, Kluge R (2017) FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas. Eur J Nucl Med Mol Imaging 44:97-110

13. Gallamini A, Barrington SF, Biggi A, et al (2014) The predictive role of interim positron

emission tomography for Hodgkin lymphoma treatment outcome is confirmed using the interpretation criteria of the Deauville five-point scale. Haematologica 99:1107–13

14. Mac Manus MP, Hicks RJ, Matthews JP, McKenzie A, Rischin D, Salminen EK, Ball DL (2003) Positron emission tomography is superior to computed tomography scanning for response-assessment after radical radiotherapy or chemoradiotherapy in patients with non-small-cell lung cancer. J Clin Oncol 21:1285–92

15. Mac Manus MP, Hicks RJ, Matthews JP, Wirth A, Rischin D, Ball DL (2005) Metabolic (FDG-PET) response after radical radiotherapy/chemoradiotherapy for non-small cell lung cancer correlates with patterns of failure. Lung Cancer 49:95–108

16. Vansteenkiste JF, Stroobants SG, De Leyn PR, Dupont PJ, Verbeken EK (1998) Potential use of FDG-PET scan after induction chemotherapy in surgically staged IIIa-N2non-small-cell lung cancer: A prospective pilot study. Ann Oncol 9(:1193–8

17. Soussan M, Chouahnia K, Maisonobe JA, Boubaya M, Eder V, Morère JF, Buvat I (2013) Prognostic implications of volume-based measurements on FDG PET/CT in stage III non-small-cell lung cancer after induction chemotherapy. Eur J Nucl Med Mol Imaging 40:668–76

18. Weber WA, Petersen V, Schmidt B, Tyndale-Hines L, Link T, Peschel C, Schwaiger M (2003) Positron emission tomography in non-small-cell lung cancer: Prediction of response to chemotherapy by quantitative assessment of glucose use. J Clin Oncol 21:2651–7

19. N. Plaxton, V. Moncayo, B. Barron RH (2014) Factors that influence standard uptake values in FDG PET/CT. J Nucl Med 55:1356–1356

20. Azmi NHM, Suppiah S, Liong CW, Noor NM, Said SM, Hanafi MH, Kaewput C, Saad FFA, Vinjamuri S (2018) Reliability of standardized uptake value normalized to lean body mass using the liver as a reference organ, in contrast-enhanced 18F-FDG PET/CT imaging. Radiat Phys Chem 147:35–39