Multi-observer concordance and accuracy of the BTS scale and other visual assessment qualitative criteria for solid pulmonary nodule (SPN) assessment with FDG PET-CT

Authors

Fatania K[1], Brown PJ[1], Xie C[2], McDermott G[3], Callister MEJ[4], Graham R[5], Subesinghe M[6,7], Gleeson FV[2], Scarsbrook AF[1,8]

Affiliations

[1]Department of Radiology, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom, UK

[2]Department of Radiology, Oxford University Hospitals Foundation Trust, Oxford, UK

[3]Department of Medical Physics, Leeds Teaching Hospitals NHS Trust, Leeds, UK

[4]Department of Respiratory Medicine, Leeds Teaching Hospitals NHS Trust, Leeds, UK

[5]Department of Radiology, Royal United Hospitals Bath NHS Foundation Trust, Bath, UK

[6]King's College London & Guy's and St. Thomas' PET Centre, St Thomas' Hospital, London, UK

[7]Department of Cancer Imaging, School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

[8]Leeds Institute of Research at St James', University of Leeds, UK

Corresponding Author

Dr Kavi Fatania, Department of Radiology, Leeds General Infirmary, Leeds, LS1 3EX, UK

kavi.fatania@nhs.net

Tel: +44(0)1132068212

Fax: +44(0)112068228

Declaration of Interest Statement

Fergus Gleeson is a share holder in Optellum

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. However, the salaries of PB and KF, and part of the salary for AS were covered by a grant from Leeds Cares.

Author Contributions

1 guarantor of integrity of the entire study – Andrew Scarsbrook

2 study concepts and design – Andrew Scarsbrook

3 literature research – Kavi Fatania, Manil Subesinghe

4 clinical studies – Kavi Fatania, Peter Brown, Cheng Xie, Garry McDermott, Matthew Callister, Richard Graham, Fergus Gleeson, Andrew Scarsbrook

5 experimental studies / data analysis – Kavi Fatania

6 statistical analysis – Kavi Fatania

7 manuscript preparation – Kavi Fatania

8 manuscript editing – Kavi, Fatania, Peter Brown, Manil Subesinghe, Andrew Scarsbrook

Abstract

Purpose

To compare the inter-observer reliability and diagnostic accuracy of the BTS scale and other visual assessment criteria in the context of FDG PET-CT evaluation of solid pulmonary nodules (SPNs).

Method

50 patients who underwent FDG PET-CT for assessment of a SPN were identified. 7 reporters with varied experience at 4 centres graded FDG uptake visually using the British Thoracic Society (BTS) 4-point scale. 5 reporters also scored SPNs according to 3- and 5-point visual assessment scales and using semi-quantitative assessment (maximum standardised uptake value - SUV_{max}). Inter-observer reliability was assessed with the intra-class correlation coefficient (ICC) and weighted Cohen's kappa (κ). Diagnostic performance was evaluated by receiver operator characteristic (ROC) analysis.

Results

Good inter-observer reliability was demonstrated with the BTS scale (ICC = 0.78, 95% CI 0.69-0.85) and 5-point scale (ICC = 0.78, 95 CI 0.68-0.86), whilst the 3-point scale demonstrated moderate reliability (ICC = 0.70, 95% CI 0.59-0.80). Almost perfect agreement was achieved between 2 consultants (κ = 0.85), and substantial agreement between 2 other consultants (κ = 0.78) using the BTS scale. ROC curves for the BTS and 5-point scales demonstrated equivalent accuracy (BTS AUC = 0.768; 5-point AUC = 0.768). SUV_{max} was no more accurate compared to the BTS scale (SUV_{max} AUC = 0.794; BTS AUC = 0.768, p = 0.43).

Conclusions

The BTS scale can be applied reliably by reporters with varied levels of PET-CT reporting experience, across different centres and has a diagnostic performance that is not surpassed by alternative scales.

Multi-observer concordance and accuracy of the BTS scale and other visual assessment qualitative criteria for solid pulmonary nodule (SPN) assessment with FDG PET-CT

Key Words

Solitary pulmonary nodule; Fluorodeoxyglucose F18; PET-CT; Reproducibility of results; Observer variation

Abbreviations

ACCP – American College of Chest Physicians

AUC – Area under the curve

BTS – British Thoracic Society

CT – Computed tomography

FDG – 2-deoxy-2-[^{18}F]fluoro-D-glucose

ICC – Intraclass correlation coefficient

IQR – Interquartile range

MBP – Mediastinal blood pool

PET – Positron emission tomography

ROC – Receiver operator curve

SPN – Solid pulmonary nodule

SUV$_{max}$ - Maximum standardised uptake value

Introduction

Risk stratification of patients found to have a solid pulmonary nodule (SPN) on imaging helps guide optimal management, allowing improved identification and treatment for malignant lesions whilst reducing intervention and harm in patients with benign disease. 2-deoxy-2-[^{18}F]fluoro-D-glucose (FDG) positron emission tomography-computed tomography (PET-CT) is widely used to non-invasively evaluate SPNs1,2 and can improve the accuracy of risk prediction models when combined with clinical risk factors3.

In UK practice, the investigation and management of patients with pulmonary nodules is based upon the 2015 British Thoracic Society (BTS) guidelines, which recommend a clinico-radiological approach to risk stratification4,5. Following the detection of a SPN on initial CT, the estimated likelihood of malignancy is determined using the Brock model6, stratifying patients into either < or > 10% risk of malignancy based upon CT findings (nodule size, count, type, location, spiculation, emphysema) and patient risk factors (age, gender, history of lung cancer). Those with >10% risk of malignancy undergo further assessment with FDG PET-CT, and risk stratification using the Herder model. The combination of SPN FDG uptake assessment and other clinico-radiological risk factors in the Herder model has been shown to improve diagnostic accuracy 3, which has been validated and confirmed in a UK population7.

The Herder model requires SPN FDG uptake to be classified according to a 4-point ordinal scale (none, faint, moderate and intense); the BTS guideline development group adapted the Herder model 4-point visual assessment scale by providing definitions for the categories of FDG uptake with reference to background uptake in the lungs and mediastinal blood pool (MBP)4,8,9. The BTS scale is the recommended method for assessment of FDG uptake in SPNs in UK practice10,11, and has been shown recently to have very good inter-observer agreement within single UK institutions 12,13. However, in order to demonstrate that this high agreement within institutions isn't due to common training methods or similar reporting techniques, it would be reassuring to reproduce these results across different institutions. Given that the BTS scale is widely used across centres in the UK, it is necessary to establish whether inter-observer agreement is of a sufficiently high standard across different UK institutions and between reporters with varying levels of PET-CT reporting experience, to confirm that the BTS scale is likely to be consistently applied nationwide. In addition, other visual assessment scales have been proposed to assess FDG uptake in SPN8, which have not been compared to the BTS scale, between reporters working across different UK institutions.

To the best of our knowledge, the BTS scale has not been assessed with regard to its inter-observer agreement between reporters working in different UK institutions, nor compared against other visual assessment scales. The aims of this study were to evaluate the inter-observer agreement across multiple reporters at 4 different UK centres and assess the relative diagnostic accuracy of 3 visual assessment scales of FDG uptake: i) BTS scale, ii) a 5-point scale modified from Fletcher et al.8, and iii) a novel 3-point visual assessment scale.

Methods

Patient selection

The reporting data set comprised initial pre-treatment FDG PET-CT scans performed in 50 patients with SPNs, who were randomly selected from an institutional database of patients at a single tertiary referral centre and who were subsequently assessed in nodule follow-up clinics between 2008 and 2013. Patients were included in this study if they had a SPN, and the diameter of their dominant SPN was between 8 and 30mm; 8mm is the minimum threshold size for resolving FDG uptake with a SPN4, and this range of nodule size reflects the standard practice of nodule assessment for UK departments7. Patients with part-solid or ground glass nodules were not included. Patients with a history of extra-pulmonary sites of malignancy and a new SPN were included as the Herder model accounts for a history of extra-pulmonary malignancy in the assessment of a SPN, and this also reflects the reality of SPN evaluation practice.

Final diagnosis was considered benign when histopathology demonstrated a benign condition, the SPN remained stable over 2 years of radiological follow-up, or the SPN spontaneously decreased or resolved without treatment. A SPN was considered malignant when histopathology confirmed primary lung cancer, there was serial interval growth of the SPN on imaging and treatment for malignancy was instigated, or the patient was known to have a histologically confirmed extra-pulmonary malignancy and new lung nodules were consistent with metastases radiologically. If patients had multiple nodules, only the largest SPN was considered for the study.

Prospective consent was obtained from all patients at the time of imaging for use of their anonymised FDG PET-CT imaging data in research and service development projects. All patients were prospectively entered into a departmental database used for retrospective identification and audit. Formal ethics committee approval was waived for this study which was considered by the institutional review board to represent evaluation of a routine clinical service.

Imaging acquisition and reconstruction

A standard protocol was used for FDG PET-CT examinations with half-body acquisition from the skull base to upper thighs. Scans prior to June 2010 were performed on a 16-slice Discovery STE PET-CT scanner (GE Healthcare, Chicago, IL, USA) and from June 2010 to December 2013 on a 64-slice Philips Gemini TF64 scanner (Philips Healthcare, Best, Netherlands). The CT component was acquired with the following settings: 140kV; 80mAs; tube rotation time 0.5 seconds per rotation; 3.75mm section thickness. Patients were asked to maintain normal shallow respiration during the CT acquisition. No iodinated contrast material was administered. Serum blood glucose was routinely checked and if blood glucose was > 10 mmol/L scanning was not performed. Patients fasted for 6 hours prior to intravenous FDG injection (dose varied according to patient body weight). All scans used iterative reconstruction (details are outlined in **Table 1**), CT for attenuation correction, applied scatter and randoms correction. Each scanner used consistent reconstruction settings, matrix and voxel size.

Image Analysis

PET-CT images for each patient were anonymised and distributed to each participating centre. Each reporter scored the FDG uptake within the dominant SPN independently, using the 3 visual assessment scales, blinded to all clinical information about the patient including eventual diagnosis. SPNs were scored using the scales outlined in **Table 2**. Each nodule was scored by visually comparing the uptake of FDG within the nodule to background tissues, including the lung parenchyma, the mediastinal blood pool (lumen of the aortic arch) and the liver, and its score assigned according to the definitions provided in **Table 2**. Examples of pulmonary nodules from each of the categories using the 5-point scale are illustrated in **Figure 1**. Mediastinal blood pool FDG uptake was determined by visually assessing uptake within the aortic arch lumen, taking care to ignore uptake in the vessel wall. Liver FDG uptake was determined by assessing the uptake within right lobe hepatic parenchyma, ignoring uptake clearly within a focal lesion (e.g. cyst), or within the vasculature.

Reporters received no additional training in the use of these visual assessment scales; the BTS scale is commonly used assessment scale in the reporting of PET-CT at each of the 4 participating centres. Reporters varied in their prior PET-CT interpretation experience: 3 ‘novice’ reporters with less than 6 months’ experience, 1 consultant radiologist who is a nuclear medicine expert with under 10 years’ experience, and 3 consultant radiologists who are nuclear medicine experts each with over 10 years’ experience. All 7 reporters assessed SPNs using the BTS scale. Due to logistical constraints, 5 out of initial 7 reporters, including 3 consultants and 2 novice reporters, also scored SPNs using the 3 and 5-point visual assessment scales and by semi-quantitative assessment (SUV_{max}) at the same time as using the BTS scale. Semi-quantitative assessment consisted of drawing a region of interest (ROI) around the SPN, and the maximum FDG uptake within this was calculated by the reporting software.

Statistical Analysis

Agreement between observers was measured using two-way random effects intraclass correlation coefficient (ICC) for multi-rater agreement and weighted Cohen's kappa (κ) for pair-wise agreement. ICC values below 0.5 indicate poor reliability, between 0.5 and 0.75 indicate moderate reliability, between 0.75 and 0.9 indicate good reliability and above 0.9 indicate excellent reliability14. Kappa values between 0.81 and 1 indicate almost perfect agreement, between 0.61 and 0.8 substantial agreement, and between 0.41 and 0.6 moderate agreement15. Diagnostic performance (i.e. discrimination of malignant from benign SPNs) of each visual assessment scale and semi-quantitative assessment with SUV_{max}, was assessed using the total area under the curve (AUC) from receiver operator characteristic (ROC) curves separately averaged across all reporters and across expert reporters only. Derivation of the averaged AUC was based on multi-rater multi-case (MRMC) statistical analysis developed by Gallas et al. and described elsewhere16, and AUCs for each assessment scale were compared using a t-test as outlined by Hillis et al.17 – this analysis was performed with the freely available software package (iMRMC: Multi-Reader, Multi-Case Analysis Methods; Version 1.2.0). Other statistical analyses were performed using SPSS (Version25; IBM, Armonk, New York, USA).

Results

Demographic data and nodule characteristics

50 patients were included in the study. Demographic information and SPN characteristics are provided in **Table 3**. The median age was 67 years (IQR 62-75 years) and 21 of the 50 patients were male (42%). 40 patients (80%) were current or former smokers and there were 37 patients (74%) with an eventual diagnosis of malignancy – 30 patients with primary lung malignancy and 7 with pulmonary metastases from an extra-pulmonary primary malignancy – the majority of patients with pulmonary metastases had metastatic colorectal carcinoma (5 patients, 10%). Median SPN diameter was 16mm (IQR 11.5-23.5mm). The mean SUV_{max} for benign SPNs was 2.5 (range 0.6-5.8), and for malignant SPNs 5.4 (range 1.2-12.4).

Interobserver agreement

Table 4 summarises the results of inter-observer agreement analysis. Inter-observer reliability for the BTS scale, for all 7 reporters including consultants and novices (ICC = 0.78, 95% CI 0.69-0.85), and between all 4 consultants (ICC = 0.77, 95% CI 0.67-0.85) was good. 5 out of 7 reporters, including 3 consultants and 2 novice reporters, also scored SPNs using the 3 and 5-point visual assessment scales and by semi-quantitative assessment (SUV_{max}). For the 5-point scale, agreement between all 5 reporters (ICC = 0.78, 95 CI 0.68-0.86), and between 3 consultants (ICC = 0.75, 95% CI 0.63-0.84) was good. For the 3-point scale, agreement between all 5 reporters (ICC = 0.70, 95% CI 0.59-0.80), and between 3 consultants (ICC = 0.64, 95% CI 0.49 0.76) was moderate.

Pair-wise analysis of agreement was performed for the BTS scale. Weighted κ demonstrated almost perfect agreement between 2 consultants, one with under (expert 1), and the other with over 10 years' experience (expert 2) (κ = 0.85), and substantial agreement between 2 consultants both with over 10 years' experience (expert 3 vs expert 4) (κ = 0.78) all working across different centres. Comparison of agreement between one consultant with over 10 years' experience with reporters of reduced experience also demonstrated substantial agreement (expert 4 vs novice 1 κ = 0.71, expert 4 vs expert 2 κ =0.75).

Diagnostic accuracy

Table 5 summarises the AUCs from ROC analysis for visual assessment scales and semi-quantitative assessment (SUV_{max}), and **Figure 2** illustrates ROC curves for each assessment method. ROCs for the BTS and 5-point scales demonstrated equivalent overall accuracy (BTS = 0.768; 5-point AUC = 0.768). The BTS scale demonstrated improved accuracy compared to the 3-point scale, although did not reach statistical significance (BTS AUC = 0.768; 3-point AUC = 0.715, p = 0.08 (Hillis, t-test)). SUV_{max} did not demonstrate statistically significant higher accuracy compared to the BTS scale (SUV_{max} AUC = 0.794; BTS AUC = 0.768, p = 0.43).

Discussion

Our study demonstrates good interobserver agreement of BTS scale, which is not improved by using a 3- or 5-point scale. The BTS scale has similar diagnostic performance across a range of reporters and sites of practice compared with other assessment methods including semi-quantitative FDG uptake measurement. The 2015 BTS guidelines for SPN evaluation advocate the use of an ordinal visual assessment scale to assess FDG uptake in SPNs on PET-CT, with the 4-point BTS scale the standard assessment scale in UK reporting practice 4,5. Murphy et al. demonstrated that the BTS scale has good inter-observer agreement within a single UK institution, using 2 different PET-CT reconstruction techniques 12 and our study further corroborates this by demonstrating good inter-observer agreement when using the BTS scale across multiple reporters from different institutions. Although the BTS scale has been advocated in national guidance, drawn together by collaborators across many institutions, this study confirms that multi-centre application of the BTS scale is reliable and extends the results of single-centre studies sharing similar conclusions 12,13. Furthermore, the study confirms that a 4-point BTS scale is not improved, with respect to its inter-observer agreement, by using a 3- or 5-point visual assessment scale. In addition, reporters of varying levels of experience showed good agreement in our study, and these results suggest that SPN risk stratification using the Herder model is likely being consistently applied across different UK centres.

Our study used visual assessment of FDG uptake within the SPN and reference background tissues to classify SPNs according to the different assessment scales (**Table 2**). In the assessment of FDG PET-CT for response assessment in Hodgkin’s and diffuse large B cell lymphoma, the 5-point scale, i.e. Deauville criteria, has demonstrated high inter-observer agreement18–20, utilising both visual assessment of FDG uptake with comparison to reference background tissues, and semi-quantitative assessment in order to confirm the results of visual assessment21. This may overcome some of the difficulties that arise from a inhomogeneous background tissue used for comparison that may lead to interobserver disagreement in visual analysis. The study by Murphy et al. demonstrated good inter-observer agreement using a similar method of visual assessment with confirmatory semi-quantitative assessment of reference background FDG uptake in the liver and blood pool. Our study shows similar results using a visual assessment of SPN FDG uptake and reference background tissue uptake, and importantly, this was observed in reporters with varying levels of experience in PET-CT reporting and across different institutions, suggesting that the BTS scale is reproducible and not due to common training in one centre alone.

The 3-point visual scale had the lowest inter-observer concordance. This could be explained by a small proportion of cases being classified on opposite ends of the 3-point scale (i.e. one reporter scored a SPN as “1” and the other as “3”), whereas they were categorized into adjacent categories for the 4-point scale (scored “2” vs “3”) or only 2 categories apart in the 5-point scale (a score of “2” vs “4”). This disagreement could not be attributed to lack of reporter experience as, even when novice reporters were excluded from analysis, 5 cases (10%) were categorised in this manner. Hence reliability was likely lower for the 3-point scale because of these cases being classified at opposite ends of the scale. It should also be noted that the reduced agreement of the 3-point scale could reflect the small sample size in our study, and that over a larger population, a difference might not have been observed. Nevertheless, the simplified 3-point scale did not perform better than the standard BTS scale recommended in the 2015 BTS guidelines.

Overall accuracy of FDG PET-CT to discriminate malignant and benign SPNs, as measured by ROC analysis, did not vary with the visual assessment scale used, and although semi-quantitative assessment of FDG uptake performed equally to visual assessment, it did not improve diagnostic accuracy to a statistically significant degree. This concurs with previous data reporting that use of semi-quantitative measurement does not improve the sensitivity of PET-CT22, but can improve its specificity23,24. Although they may not have played a significant role in our study, in general there are several factors that can limit the use of a semi-quantitative measure for distinguishing malignant and benign SPNs. First, technical factors can limit the standardisation of SUV values across different scanner and sites where scan technique, for example reconstruction algorithms, may vary and therefore so too will the SUV measurements25. All the images used in this study were acquired in a single institution. Using an alternative reconstruction algorithm has recently been shown to increase the Herder score for SPNs, although not the overall diagnostic performance of the Herder scale, for example 12. Second, studies utilising semi-quantitative measures typically use a single cut-off value to distinguish benign and malignant nodules26, and typically do not include a validation cohort to test their cut-off values9,27, whereas the use of visual ordinal scales can reflect increasing likelihood that a nodule is malignant and overcome the difficulties of semiquantitative measurement8. Lastly, the calculated SUV can be erroneous due to tracer extravasation or inaccurate patient weight.

The diagnostic accuracy of both visual assessment scales and semi-quantitative measurements were lower in this study than previously reported by others2,9,28. This may be explained by the high proportion of malignant SPNs included in this patient cohort, which might have influenced test sensitivity and specificity29. Our results are similar to those of Lopez et al. who also had a high prevalence of malignant nodules in their study sample23 and to Murphy et al. whose prevalence of malignancy was 77%12. The high proportion of current or former smokers in our patient cohort is also likely to have influenced the AUC, as it is known that in higher risk patients, FDG PET-CT has reduced specificity9. Finally, the mean SUV$_{max}$ for benign SPNs in the study was 2.5, which in other studies27,30 is taken as the threshold for assigning a nodule as malignant on PET-CT, suggesting that our sample may have over-represented benign SPNs (i.e. inflammatory or infective SPNs) with 'false' positive FDG uptake31 compared with other studies. This will have further reduced the specificity of assessment. The accuracy of visual assessment might have been improved by using semi-quantitative assessment of uptake in reference tissues to confirm the results of visual assessment, as used in Deauville criteria21 and by Murphy et al12.

The study had a number of important limitations. First, 50 patients is a relatively small sample size, and it is possible that a larger cohort may have revealed differences in accuracy and/or reliability between the BTS and 5-point scales. Second, not all diagnoses were confirmed histologically, and therefore it is possible that this introduced inaccuracy in the classification of a SPN being definitely malignant or benign, again which would affect the overall diagnostic accuracy. However, each scale would be similarly affected, and this should not limit the comparison between them. Furthermore, these criteria reflect the reality of clinical practice, when treatment decisions are not always based on histological diagnosis. The images used in assessment were acquired on different scanners, using different imaging conditions which introduces a potential source of variation in the image quality, however this should not have a strong effect on the comparative assessment of different assessment methods. Lastly, this was a retrospective analysis on non-consecutive patients which is potentially a source of bias, however this would have affected each assessment scale equally and is unlikely to affect our conclusions.

Conclusion

Our study confirms recent single-centre experiences and extends this to demonstrate that the BTS scale can be applied consistently in the assessment of SPNs by observers working at different centres and by individuals with limited prior PET-CT interpretation experience. The BTS scale is advocated in national guidance for evaluation of SPN's and although it would be expected that the scale is easily reproducible across multiple institutions, our study confirms that this is the case. The BTS scale, which is being increasingly used as part of risk stratification of SPNs has an accuracy which is not surpassed by alternative visual or semi-quantitative assessment scales.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent

Informed consent was obtained from all individual participants included in the study.

References

1. Gould MK, Maclean CC, Kuschner WG, Rydzak CE, Owens DK. Accuracy of Positron Emission Tomography for Diagnosis of Pulmonary Nodules and Mass Lesions. JAMA 2001;**285**(7):914. https://doi.org/10.1001/jama.285.7.914.
2. Ruilong Z, Daohai X, Li G, Xiaohong W, Chunjie W, Lei T. Diagnostic value of 18F-FDG-PET/CT for the evaluation of solitary pulmonary nodules. Nucl Med Commun 2017;**38**(1):67–75. https://doi.org/10.1097/MNM.0000000000000605.
3. Herder GJ, van Tinteren H, Golding RP, *et al.* Clinical Prediction Model To Characterize Pulmonary Nodules. Chest 2005;**128**(4):2490–6. https://doi.org/10.1378/chest.128.4.2490.
4. Callister MEJ, Baldwin DR, Akram AR, *et al.* British Thoracic Society guidelines for the investigation and management of pulmonary nodules: accredited by NICE. Thorax 2015;**70**(Suppl 2):ii1–54. https://doi.org/10.1136/thoraxjnl-2015-207168.
5. Graham RNJ, Baldwin DR, Callister MEJ, Gleeson F V. Return of the pulmonary nodule: the radiologist’s key role in implementing the 2015 BTS guidelines on the investigation and management of pulmonary nodules. Br J Radiol 2016;**89**(1059):20150776. https://doi.org/10.1259/bjr.20150776.
6. McWilliams A, Tammemagi MC, Mayo JR, *et al.* Probability of Cancer in Pulmonary Nodules Detected on First Screening CT. N Engl J Med 2013;**369**(10):910–9. https://doi.org/10.1056/NEJMoa1214726.
7. Al-Ameri A, Malhotra P, Thygesen H, *et al.* Risk of malignancy in pulmonary nodules: A validation study of four prediction models. Lung Cancer 2015;**89**(1):27–30. https://doi.org/10.1016/j.lungcan.2015.03.018.

8. Fletcher JW, Kymes SM, Gould M, *et al.* A Comparison of the Diagnostic Accuracy of 18F-FDG PET and CT in the Characterization of Solitary Pulmonary Nodules. J Nucl Med 2008;**49**(2):179–85. https://doi.org/10.2967/jnumed.107.044990.

9. Evangelista L, Cuocolo A, Pace L, *et al.* Performance of FDG-PET/CT in solitary pulmonary nodule based on pre-test likelihood of malignancy: results from the ITALIAN retrospective multicenter trial. Eur J Nucl Med Mol Imaging 2018;**45**(11):1898–907. https://doi.org/10.1007/s00259-018-4016-1.

10. Callister MEJ, Baldwin DR. How should pulmonary nodules be optimally investigated and managed? Lung Cancer 2016;**91**:48–55. https://doi.org/10.1016/j.lungcan.2015.10.018.

11. Baldwin D, Callister M, Akram A, *et al.* British Thoracic Society quality standards for the investigation and management of pulmonary nodules. BMJ Open Respir Res 2018;**5**(1):e000273. https://doi.org/10.1136/bmjresp-2017-000273.

12. Murphy D, Royle L, Chalampalakis Z, *et al.* The effect of a novel Bayesian penalised likelihood PET reconstruction algorithm on the assessment of malignancy risk in solitary pulmonary nodules according to the British Thoracic Society guidelines. Eur J Radiol 2019;**117**:149–55. https://doi.org/10.1016/j.ejrad.2019.06.005.

13. Ordidge KL, Gandy N, Arshad MA, *et al.* Interobserver agreement of the visual Herder scale for the assessment of solitary pulmonary nodules on 18F Fluorodeoxyglucose PET/computed tomography. Nucl Med Commun 2020;**41**(3):235–40. https://doi.org/10.1097/mnm.0000000000001146.

14. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;**15**(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

15. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics 1977;**33**(1):159. https://doi.org/10.2307/2529310.

16. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. Acad Radiol 2006;**13**(3):353–62. https://doi.org/10.1016/j.acra.2005.11.030.

17. Hillis SL, Berbaum KS, Metz CE. Recent Developments in the Dorfman-Berbaum-Metz Procedure for Multireader ROC Study Analysis. Acad Radiol 2008;**15**(5):647–61. https://doi.org/10.1016/j.acra.2007.12.015.

18. Barrington SF, Qian W, Somer EJ, *et al.* Concordance between four European centres of PET reporting criteria designed for use in multicentre trials in Hodgkin lymphoma. Eur J Nucl Med Mol Imaging 2010;**37**(10):1824–33. https://doi.org/10.1007/s00259-010-1490-5.

19. Biggi A, Gallamini A, Chauvie S, *et al.* International Validation Study for Interim PET in ABVD-Treated, Advanced-Stage Hodgkin Lymphoma: Interpretation Criteria and Concordance Rate Among Reviewers. J Nucl Med 2013;**54**(5):683–90. https://doi.org/10.2967/jnumed.112.110890.

20. Burggraaff CN, Cornelisse AC, Hoekstra OS, *et al.* Interobserver Agreement of Interim and End-of-Treatment 18 F-FDG PET/CT in Diffuse Large B-Cell Lymphoma: Impact on Clinical Practice and Trials. J Nucl Med 2018;**59**(12):1831–6. https://doi.org/10.2967/jnumed.118.210807.

21. Barrington SF, Kluge R. FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas. Eur J Nucl Med Mol Imaging 2017;**44**:97–110. https://doi.org/10.1007/s00259-017-3690-8.

22. Nomori H, Watanabe K, Ohtsuka T, Naruke T, Suemasu K, Uno K. Visual and Semiquantitative Analyses for F-18 Fluorodeoxyglucose PET Scanning in Pulmonary

Nodules 1 cm to 3 cm in Size. Ann Thorac Surg 2005;**79**(3):984–8. https://doi.org/10.1016/j.athoracsur.2004.07.072.

23. López OVG, María A, Vicente G, *et al.* F-FDG-PET/CT in the assessment of pulmonary solitary nodules: comparison of different analysis methods and risk variables in the prediction of malignancy. Transl Lung Cancer Res 2015;**4**(3):228–35. https://doi.org/10.3978/j.issn.2218-6751.2015.05.07.

24. Hashimoto Y, Tsujikawa T, Kondo C, *et al.* Accuracy of PET for diagnosis of solid pulmonary lesions with 18F-FDG uptake below the standardized uptake value of 2.5. J Nucl Med 2006;**47**(3):426–31.

25. Boellaard R. Standards for PET Image Acquisition and Quantitative Data Analysis. J Nucl Med 2009;**50**(Suppl_1):11S-20S. https://doi.org/10.2967/jnumed.108.057182.

26. Gould MK, Donington J, Lynch WR, *et al.* Evaluation of Individuals With Pulmonary Nodules: When Is It Lung Cancer? Chest 2013;**143**(5):e93S-e120S. https://doi.org/10.1378/chest.12-2351.

27. Li S, Zhao B, Wang X, *et al.* Overestimated value of 18 F-FDG PET/CT to diagnose pulmonary nodules: Analysis of 298 patients. Clin Radiol 2014;**69**(8):352–7. https://doi.org/10.1016/j.crad.2014.04.007.

28. Cronin P, Dwamena BA, Kelly AM, Carlos RC. Solitary Pulmonary Nodules: Meta-analytic Comparison of Cross-sectional Imaging Modalities for Diagnosis of Malignancy. Radiology 2008;**246**(3):772–82. https://doi.org/10.1148/radiol.2463062148.

29. Leeflang MMG, Rutjes AWS, Reitsma JB, Hooft L, Bossuyt PMM. Variation of a test's sensitivity and specificity with disease prevalence. Can Med Assoc J 2013;**185**(11):E537–44. https://doi.org/10.1503/cmaj.121286.

30. Orlacchio A, Schillaci O, Antonelli L, *et al.* Solitary pulmonary nodules: morphological and metabolic characterisation by FDG-PET-MDCT. Radiol Med 2007;**112**(2):157–73. https://doi.org/10.1007/s11547-007-0132-x.

31. Rosenbaum SJ, Lind T, Antoch G, Bockisch A. False-Positive FDG PET Uptake−the Role of PET/CT. Eur Radiol 2006;**16**(5):1054–65. https://doi.org/10.1007/s00330-005-0088-y.

Figures and tables

Table 1 - Reconstruction parameters for each scanner

Table 2 - Visual assessment scale scoring criteria

Table 3 – Demographic data and SPN characteristics (n=50)

Table 4 – Inter-observer agreement for visual assessment scales

Table 5 - Accuracy of visual assessment scales and semiquantitative assessment

Figure 1 – Examples of pulmonary nodules demonstrating increasing FDG uptake

Caption for Figure 1:

Maximum intensity projection (MIP) image from 5 patients with SPN that demonstrate increasing FDG uptake (from right to left), and illustrate examples of each category using the 5-point visual assessment scale. From the right-hand image, an example of no uptake, through to the left-hand image showing uptake above that of the liver. Black circles indicate the location of the SPN being assessed. MBP = mediastinal blood pool.

Figure 2 – Receiver operator curves for visual assessment scales and semiquantitative assessment

Caption for Figure 2:

4 receiver-operator curves demonstrating similar diagnostic performance For visual uptake scales and semiquantitative assessment compared to the BTS scale.

Table 1 - Reconstruction parameters for each scanner

Scanner	**Reconstruction**	**Scatter correction**	**Randoms correction**	**Matrix**	**Voxel size (x,y,z mm)**
GE Healthcare STE	OSEM	Convolution subtraction	Singles	128	4.7 x 4.7 x 3.3
Philips Gemini TF64	BLOB-OS-TF	SS-Simul	DLYD	144 or 169	4.0 x 4.0 x 4.0

Key:

OSEM – Ordered subsets expectation maximisation

Uptake	3-point scale	BTS scale	5-point scale
Indiscernible from background lung	1	1	1
Greater than lung but less MBP		2	2
Equal to MBP	2		3
Greater than MBP but less than liver	3	3	4
Greater than liver		4	5

MBP – mediastinal blood pool

BLOB-OS-TF – Spherically symmetric basis function ordered subset algorithm

DLYD – delayed event subtraction

Table 2 - Visual assessment scale scoring criteria

Table 3 – Demographic data and nodule characteristics (n=50)

Demographic	Value
Median age, years (IQR)	67 (62-75)
Male gender (%)	21 (42%)
Smoking status (%)	
Current or former smoker	40 (80%)
Never smoked	7 (14%)
Smoking status undocumented	3 (6%)
Diagnosis (%)	
Primary lung cancer	30 (60%)
Metastases from extra-pulmonary primary malignancy	7 (14%)
Colorectal adenocarcinoma	5 (10%)
Cervical squamous cell carcinoma	1 (2%)
Pancreatic large cell carcinoma	1 (2%)
Benign nodule	13 (26%)
Median nodule diameter, mm (IQR)	16 (11.5 – 23.5)

IQR = interquartile range

Table 4 – Inter-observer agreement for visual assessment scales

Visual assessment scale	Agreement: All observers ICC (95% CI)	Agreement: Expert observers ICC (95% CI)
3-point scale	0.70 (0.59 - 0.80)	0.64 (0.49 - 0.76)
BTS scale	0.78 (0.69 - 0.85)	0.77 (0.67 - 0.85)
5-point scale	0.78 (0.68 - 0.86)	0.75 (0.63 - 0.84)

ICC = 2-way random effects intra-class correlation coefficient

Table 5 - Accuracy of visual assessment scales and semiquantitative assessment

Assessment method	Area under ROC	*p* value[*] (versus 4-point BTS scale)
3-point scale	0.715	0.08
BTS scale	0.768	NA
5-point scale	0.768	NA
SUV_{max}	0.794	0.43

* t-test – as outlined by Hillis et al.

NA – not applicable

Figure 1

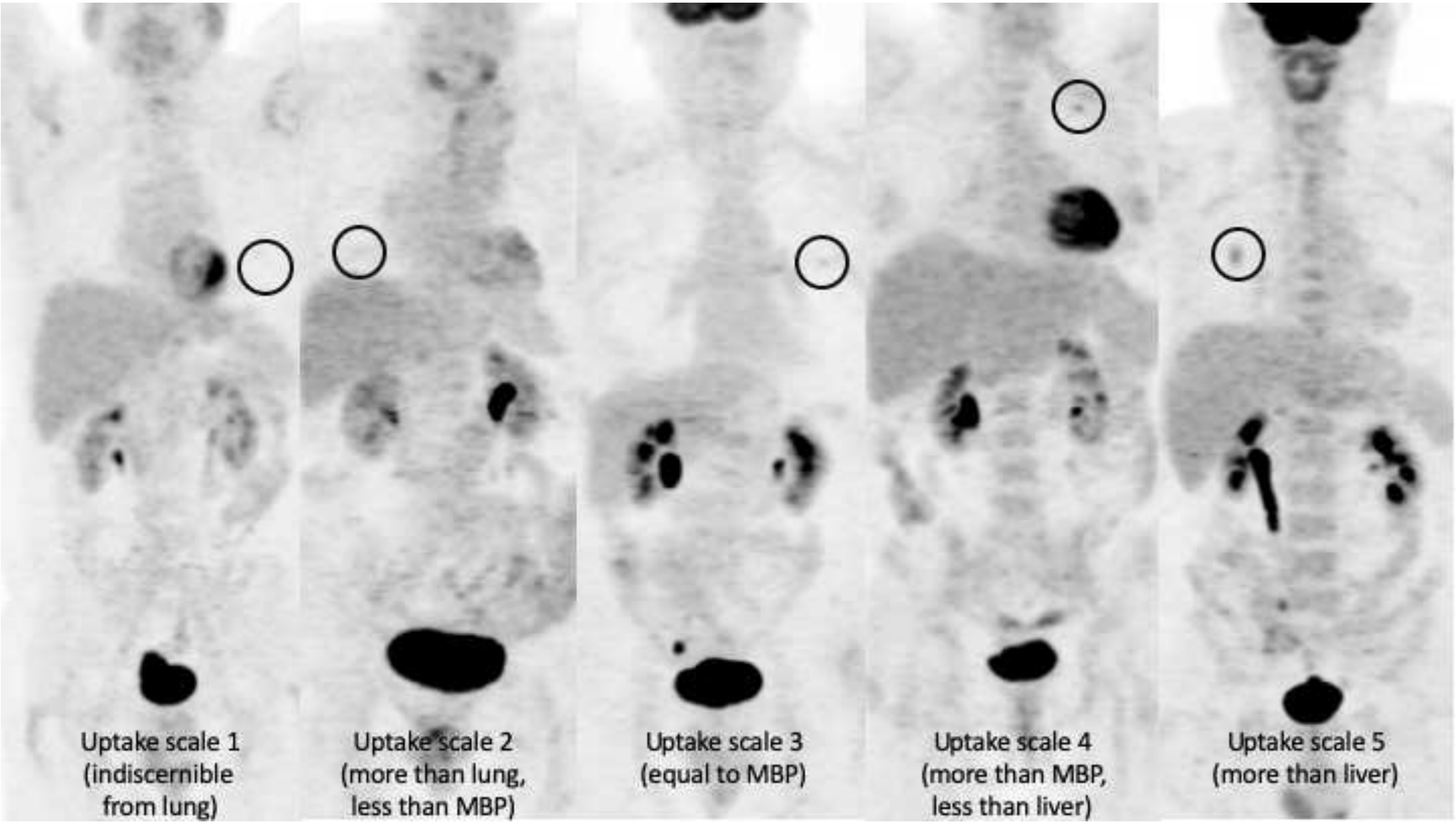
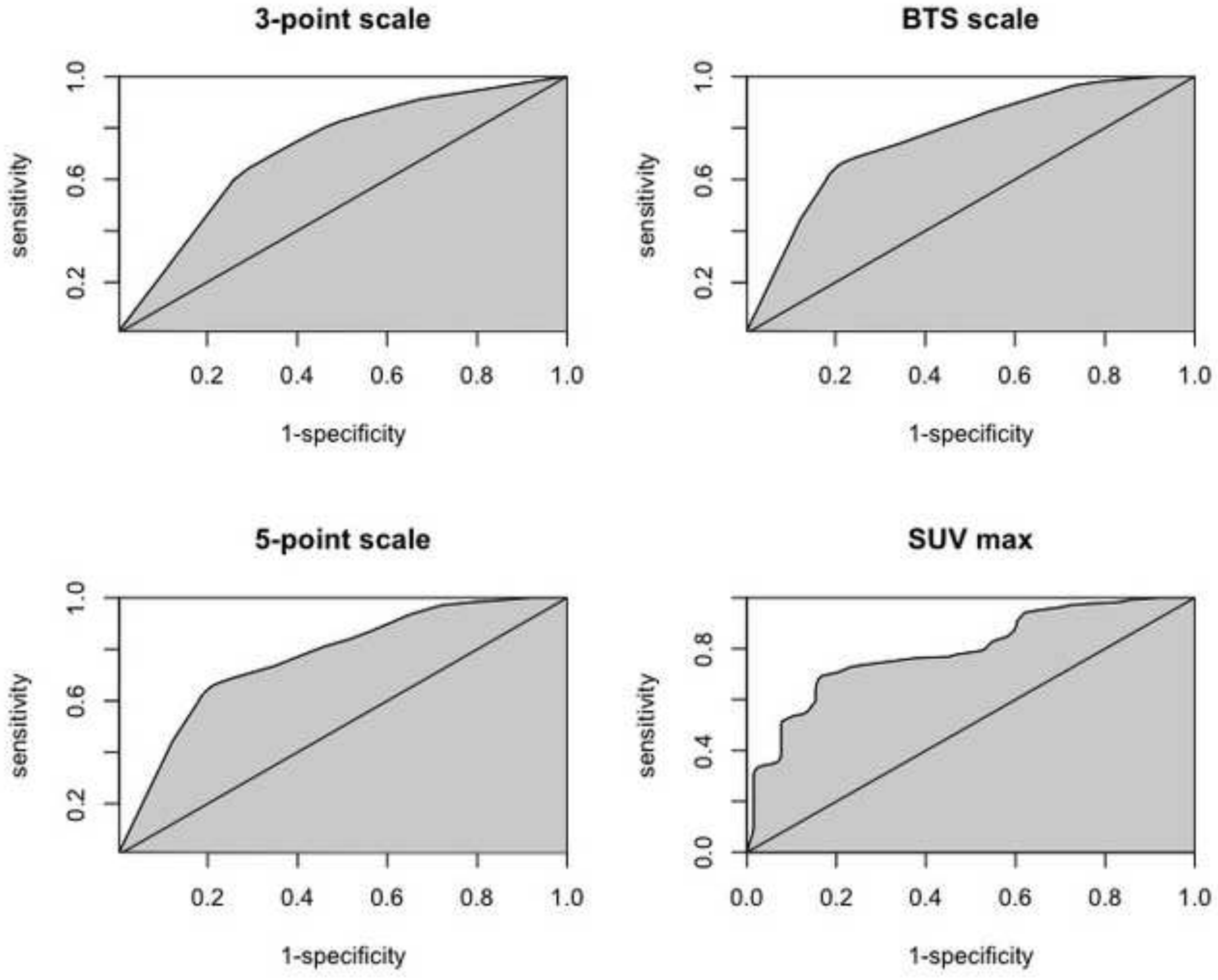

Uptake scale 1
(indiscernible
from lung)
Uptake scale 2
(more than lung,
less than MBP)
Uptake scale 3
(equal to MBP)
Uptake scale 4
(more than MBP,
less than liver)
Uptake scale 5
(more than liver)

Figure 2


3-point scale
BTS scale
5-point scale
SUV max
sensitivity
1-specificity
0.0
0.2
0.4
0.6
0.8
1.0

Highlights

- British Thoracic Society scale of FDG uptake has good inter-observer agreement.
- British Thoracic Society scale is as reliable as 3 and 5 point visual scales.
- Visual assessment showed good agreement between reporters across institutions.
- Semi-quantitative assessment did not improve the diagnostic accuracy.