Received: 28 January 2022 | Revised: 15 June 2022 | Accepted: 22 June 2022

DOI: 10.1111/cpf.12773

ORIGINAL ARTICLE

Clinical Physiology and Functional Imaging

Automated classification of PET-CT lesions in lung cancer: An independent validation study

Pablo Borrelli[1] | José Luis Loaiza Góngora[1] | Reza Kaboteh[1] | Olof Enqvist[2] | Lars Edenbrandt[1,3]

[1]Department of Clinical Physiology, Sahlgrenska University Hospital, Region Västra Götaland, Gothenburg, Sweden

[2]Eigenvision AB, Malmö, Sweden

[3]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Correspondence

Lars Edenbrandt, Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, 413 45 Gothenburg, Sweden.
Email: lars.edenbrandt@gu.se

Abstract

Introduction: Recently, a tool called the positron emission tomography (PET)-assisted reporting system (PARS) was developed and presented to classify lesions in PET/computed tomography (CT) studies in patients with lung cancer or lymphoma. The aim of this study was to validate PARS with an independent group of lung-cancer patients using manual lesion segmentations as a reference standard, as well as to evaluate the association between PARS-based measurements and overall survival (OS).

Methods: This study retrospectively included 115 patients who had undergone clinically indicated (18F)-fluorodeoxyglucose (FDG) PET/CT due to suspected or known lung cancer. The patients had a median age of 66 years (interquartile range [IQR]: 61–72 years). Segmentations were made manually by visual inspection in a consensus reading by two nuclear medicine specialists and used as a reference. The research prototype PARS was used to automatically analyse all the PET/CT studies. The PET foci classified as suspicious by PARS were compared with the manual segmentations. No manual corrections were applied. Total lesion glycolysis (TLG) was calculated based on the manual and PARS-based lung-tumour segmentations. Associations between TLG and OS were investigated using Cox analysis.

Results: PARS showed sensitivities for lung tumours of 55.6% per lesion and 80.2% per patient. Both manual and PARS TLG were significantly associated with OS.

Conclusion: Automatically calculated TLG by PARS contains prognostic information comparable to manually measured TLG in patients with known or suspected lung cancer. The low sensitivity at both the lesion and patient levels makes the present version of PARS less useful to support clinical reading, reporting and staging.

KEYWORDS

artificial intelligence, computer-based methods, image analysis, quantification, survival analysis

Pablo Borrelli and José Luis Loaiza Góngora shared first authorship.

1 | INTRODUCTION

The clinical demand for medical imaging services is increasing rapidly (Smith-Bindman et al., 2008, 2019). One example is the use of positron emission tomography with computed tomography (PET/CT) for the characterization of lung lesions (Groheux et al., 2016). However, the increased number of studies to be interpreted is not matched by an increased supply of qualified radiologists and nuclear medicine specialists. In addition, manually reading these types of three-dimensional images is time-consuming, and the reports usually lack quantification and show poor reproducibility (Heye et al., 2013). Therefore, the true clinical potential of the diagnostic images is not realized.

Artificial Intelligence (AI) offers new opportunities to support radiologists and nuclear medicine specialists to increase output without compromising quality (Rodriguez-Ruiz et al., 2019). AI tools could decrease reading times, improve reproducibility and reliably quantify cancer-related PET-tracer activity (Hosny et al., 2018). Recently, Sibille et al. presented an AI tool to classify fluorine 18 (18F)-fluorodeoxyglucose (FDG) avid lesions in whole-body PET/CT studies for patients with lung cancer or lymphoma (Sibille et al., 2020). The tool showed promising accuracy, but the authors pointed out that further studies are required to develop the approach into a clinical tool that supports clinical reading, automated reporting and staging. The tool presented by Sibille et al. is incorporated in a research prototype called the PET-assisted reporting system (PARS), which is available for research purposes. The same research group has been involved in two studies evaluating PARS (Capobianco et al., 2021; Pinochet et al., 2021). In addition, Weber et al. performed an independent evaluation of PARS in patients with breast cancer in cases that had not been used for the training of the system (Weber et al., 2021).

Bluemke et al. pointed out in an Editorial in Radiology that AI tools need independent validation and should therefore be publicly available so that performance claims can be verified (Bluemke et al., 2020). Thus, the aim of this study was to compare the automatic performance (sensitivity and specificity) of PARS in an independent group of patients with known or suspected lung cancer using manual lesion segmentations by two nuclear medicine specialists as a reference standard. A second aim was to assess whether total lesion glycolysis (TLG) calculated from PARS-based segmentations of lung tumours shows comparable prognostic value to manually measured TLG.

2 | METHODS

2.1 | Patients

This retrospective analysis included 115 consecutive patients who underwent clinically indicated FDG PET/CT due to suspected lung cancer (*n* = 63) or for the management of known lung cancer (*n* = 52) between April 2008 and December 2010. We used this group of patients in a previous study to train and evaluate an AI tool for the detection of lung tumours (Borrelli et al., 2021). The patient group consisted of 60 females and 55 males with a median age of 66 years (IQR: 61–72 years). Clinical information and survival data were collected from the local medical records and radiology information system up until April 2021. A total of 84 patients died during the follow-up period, and the median survival time from the PET/CT study was 2.3 years (IQR: 1.2–5.5 years). The group of 31 patients who were still alive had a median follow-up time from the PET/CT study of 11.9 years (IQR: 11.6–12.2 years). This study was approved by the Regional Ethical Review Board at the University of Gothenburg, Sweden (#295-08) and was performed in accordance with the Declaration of Helsinki. All patients provided written informed consent.

The patients were injected with 4 MBq/kg (maximum of 400 Mbq) of FDG, fasted for at least 4–6 h before the injection, and had adequate glucose levels before injection. The accumulation time was 60 min.

PET/CT scans were obtained using an integrated PET/CT system (Siemens Biograph 64 Truepoint). Images were acquired with 3 min per bed position from the base of the skull to the mid-thigh. PET images were reconstructed with a slice thickness of 3 mm using an iterative ordered subset expectation maximization 3D algorithm (four iterations and eight subsets). The matrix size was 168 × 168, and CT-based attenuation and scatter corrections were applied. A low-dose CT scan (64-slice helical, 120 kV, 30 mAs, 512 × 512 matrix) was obtained for the same part of the patient as the PET scan. The CT was reconstructed using a filtered back projection algorithm with slice thickness and spacing matching the PET scan.

2.2 | Manual segmentations

Manual segmentations in the PET images were performed by two nuclear medicine specialists with >6 years of PET/CT experience and used as a reference. No clinical data, results from other imaging modalities or survival data were available to the readers. The segmentations were made manually by visual inspection in a consensus reading. A suspect lesion was considered when there was a morphological lesion in the CT that were FDG-avid (SUVmax lesion > SUVmax blood pool). Each lesion was classified as a lung tumour or lymph node. A cloud-based annotation tool (RECOMIA, https://www.recomia.org) was used to segment the PET/CT studies (Trägårdh et al., 2020). TLG was calculated based on the manual segmentations of lung tumours.

2.3 | Research prototype

The PARS research prototype (version 3.0, Siemens Medical Solutions USA, Inc.) was used to automatically analyse all the PET/CT studies. The analysis included segmentation of the liver as reference region, segmentation of PET foci using a thresholding algorithm and classification of anatomical location and suspiciousness for all detected PET foci using a convolutional neural network. The PET foci classified as suspicious (physiological uptake=FALSE) by PARS were marked in red and physiological uptake in green. An example of the PARS display is shown in Figure 1. The suspicious foci by PARS were compared with the manual segmentations. No preprocessing of the PET/CT studies or manual corrections of the PARS analysis were applied. TLG was calculated based on the segmentations of lung tumours by PARS and manual segmentations.

2.4 | Statistical methods

Sensitivity, specificity, accuracy and positive predictive value were calculated at the per-patient and lesion levels. Associations between TLG and overall survival (OS) were investigated using a univariate Cox proportional hazards regression model. OS was calculated from the date of the PET/CT study to the date of death or last follow-up. Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated. The TLG measurements had a skewed distribution and were log10 transformed after adding 1.0 to handle zeros.

Bland–Altman plot was used to compare PARS and manual TLG. Bivariate proportional regression Cox analyses were performed to evaluate the association between manual and PARS TLGs and OS. OS curves were produced according to the Kaplan–Meier method. The curves were produced after dividing the patients into three groups based on whether their TLG values were higher or lower than the 1/3 and 2/3 quantiles. Only two groups were used for the smaller subsets of patients with diagnosed and suspected cancer. The statistical analysis was performed in R (version 4.0.3).

3 | RESULTS

PARS showed a per-lesion sensitivity of 55.6% (79/142) for lung tumours and 35.3% (63/185) for lymph nodes using the manual segmentations by the two nuclear medicine specialists as the reference standard (Table 1). The positive predictive value was high for lung tumour lesions (94.0%) and low for lymph nodes (70.0%).

The per-patient sensitivity and specificity for lung tumours were 80.2% (69/86) and 100% (29/29), respectively (Table 2). The sensitivity and specificity for lymph nodes were 70.2% (40/57) and 81.0% (47/58), respectively. The positive predictive value was 100% (69/69) for lung tumours and 78.4% (40/51) for lymph nodes. The negative predictive value was 63% (29/46) for lung tumours and 73.4% (47/64) for lymph nodes. PARS did not detect any suspicious PET focus in 7 of the 17 false-negative studies. In the remaining 10 studies, no lung tumour was detected, but other suspicious PET foci classified as lymph nodes or metastases in the bone, adrenal gland or small intestine were detected.

Manual and PARS segmentations of lung tumours were used to compute TLG for each patient. Figure 2 shows a Bland–Altman plot comparing PARS and manual TLG. The mean difference between them was −22.6 with a standard deviation of 237. Both the manual and PARS TLGs were significantly associated with OS in uni- and bivariate proportional regression Cox analyses (Table 3).

TABLE 1 Per-lesion results of PARS

	PARS positive	PARS negative	Total
Lung tumours			
Manual positive	79	63	142
Manual negative	5	-	
Total	84		
Lymph nodes			
Manual positive	63	122	185
Manual negative	27		
Total	90		

Abbreviation: PARS, PET-assisted reporting system.

TABLE 2 Per-patient results of PARS

	PARS positive	PARS negative	Total
Lung tumours			
Manual positive	69	17	86
Manual negative	0	29	29
Total	69	46	115
Lymph nodes			
Manual positive	40	17	57
Manual negative	11	47	58
Total	51	64	115

Abbreviation: PARS, PET-assisted reporting system.

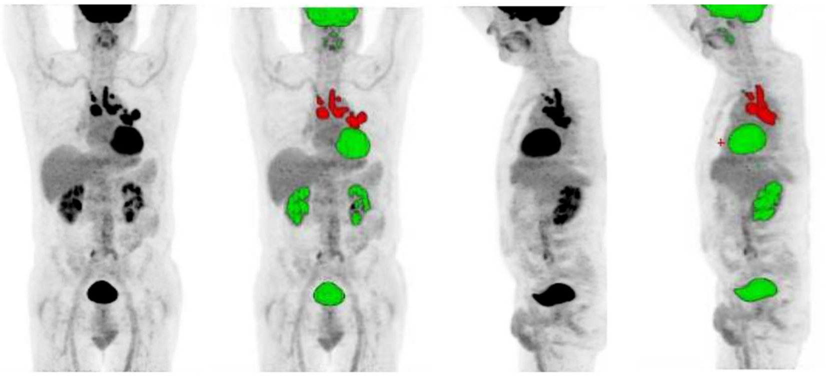

FIGURE 1 Maximum intensity projections of a patient with positron emission tomography (PET) regions classified as suspicious by PET-assisted reporting system (PARS) in red and physiological uptake in green

The patients were divided into three equally sized quantiles based on their PARS TLG. The Kaplan–Meier curves in Figure 3 show the survival probabilities for these groups. The median survival times for the group with the highest, middle and lowest PARS TLG were 1.5 years, 4.2 years and not reached after 5 years of follow-up. The corresponding Kaplan–Meier curves for manual TLG showed median survival times for the three groups of 1.4 years, 4.2 years and not reached after 5 years of follow-up.

In a separate analysis of the patients with known lung cancer at the time of the PET/CT study (*n* = 52), patients with PARS TLG that were greater than the median value had a shorter median survival time (1.6 years) than the patients with values that were less than the median (not reached after 5 years of follow-up) (Figure 4). Identical results were obtained using manual TLG. Performing the same analysis for the patients with suspected lung cancer at the time of the PET/CT study (*n* = 63) showed median survival times for the two groups of 2.8 years and not reached after 5 years of follow-up based on PARS TLG. The corresponding results for manual TLG was 2.3 years and not reached after 5 years of follow-up, respectively.

4 | DISCUSSION

This independent validation of the research prototype PARS showed a per-patient sensitivity of 80.2% and specificity of 100% for lung tumours. These results are close to those presented in the original article by Sibille et al. (sensitivity 87.1%; specificity 99.0%) (Sibille et al., 2020). We used the same type of standard reference (manual segmentations by experts) in our study group, which was more than twice as large as the test set of Sibille et al. (*n* = 115 vs. 59). One difference between the studies was that we validated the complete analysis of PARS, that is, detection and classification of PET foci, whereas Sibille et al. only presented results for the classification of PET foci based on a convolutional neural network. The detection part of PARS, which is based on a fixed thresholding algorithm, was not included in that study. Sibille et al. pointed out that the data that they used to train, validate and test PARS were from a single site and may therefore not generalize well to sites with different imaging protocols, cameras and patient characteristics. To our knowledge, this is the first independent validation of PARS in lung cancer patients to confirm the results of the original study.

Recently, two other validation studies of PARS have been presented. Weber et al. (2021) applied PARS to PET/CT studies from patients with breast cancer in cases that were not used in the training of PARS. They found a per-patient accuracy of PARS in detecting lesions of 72%, which is close to the results for lymph nodes in our study (75.7%). The accuracy for lung tumours was higher in our study, which could at least partly be explained by PARS being trained on studies from patients with lung cancer or lymphoma.

Capobianco et al. and Pinochet et al. validated PARS in patients with lymphoma (Capobianco et al., 2021; Pinochet et al., 2021). They both found lower PARS-based total metabolic tumour volumes (TMTVs) than the corresponding manual TMTVs. The differences between PARS and

TABLE 3 Association between TLG, age and overall survival

Variables	HR (95% CI)	*p* value
Univariate		
Manual TLG	1.99 (1.54–2.57)	<0.001
PARS TLG	1.90 (1.50–2.41)	<0.001
Bivariate		
Manual TLG	1.99 (1.53–2.58)	<0.001
Age	1.01 (0.984–1.04)	0.42
PARS TLG	1.89 (1.49–2.40)	<0.001
Age	1.01 (0.979–1.03)	0.65

Abbreviations: HR, hazard ratio; CI, confidence interval; PARS, PET-assisted reporting system; TLG, total lesion glycolysis.

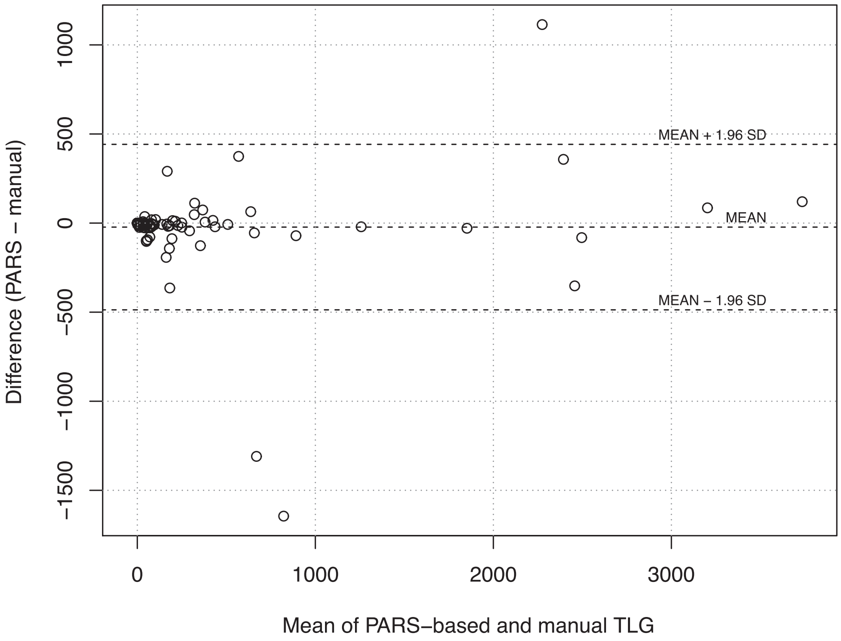

FIGURE 2 Bland–Altman plot comparing manual and PET-assisted reporting system (PARS) total lesion glycolysis (TLG)

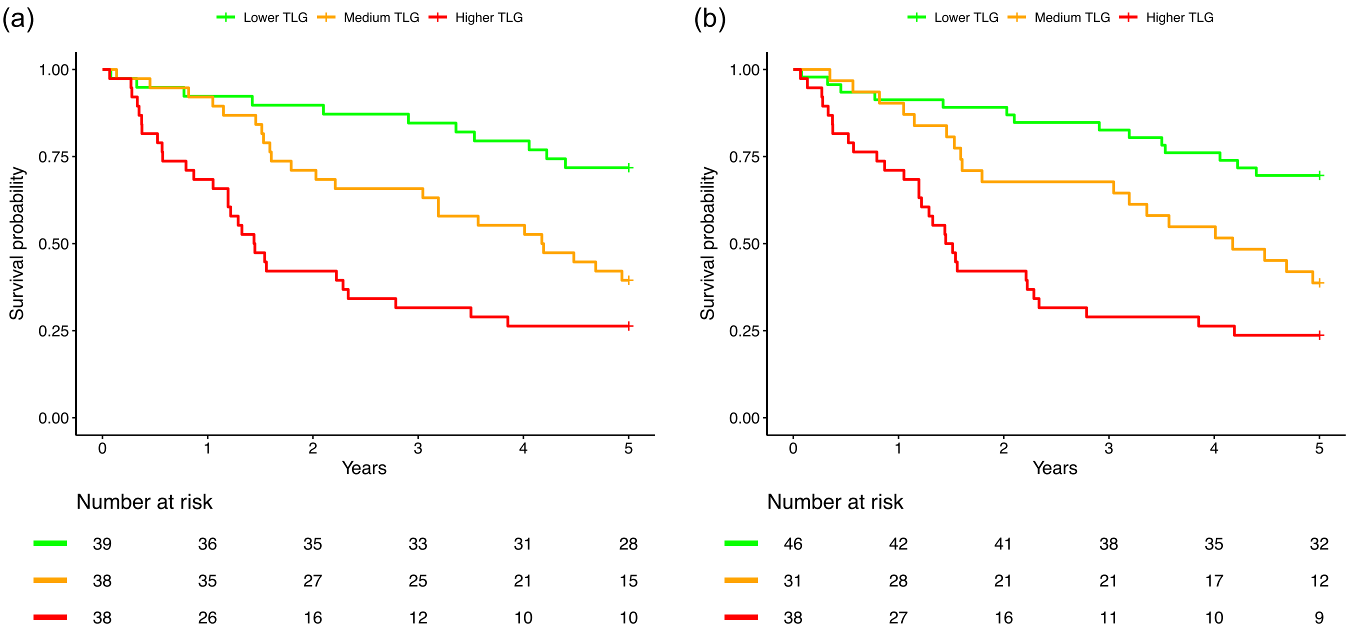

FIGURE 3 Kaplan–Meier curves showing survival probabilities for patients divided into three similarly sized groups based on whether their manual total lesion glycolysis (TLG) (left) and PET-assisted reporting system (PARS) TLG (right) values are smaller or larger than the 1/3-quantile and the 2/3-quantile ($n = 115$). The TLG ranges for the groups were 0–7, 7–103 and 103–3675 for the groups based on manual TLG and 0–0, 3–77 and 77–3795 for the groups based on PARS TLG.

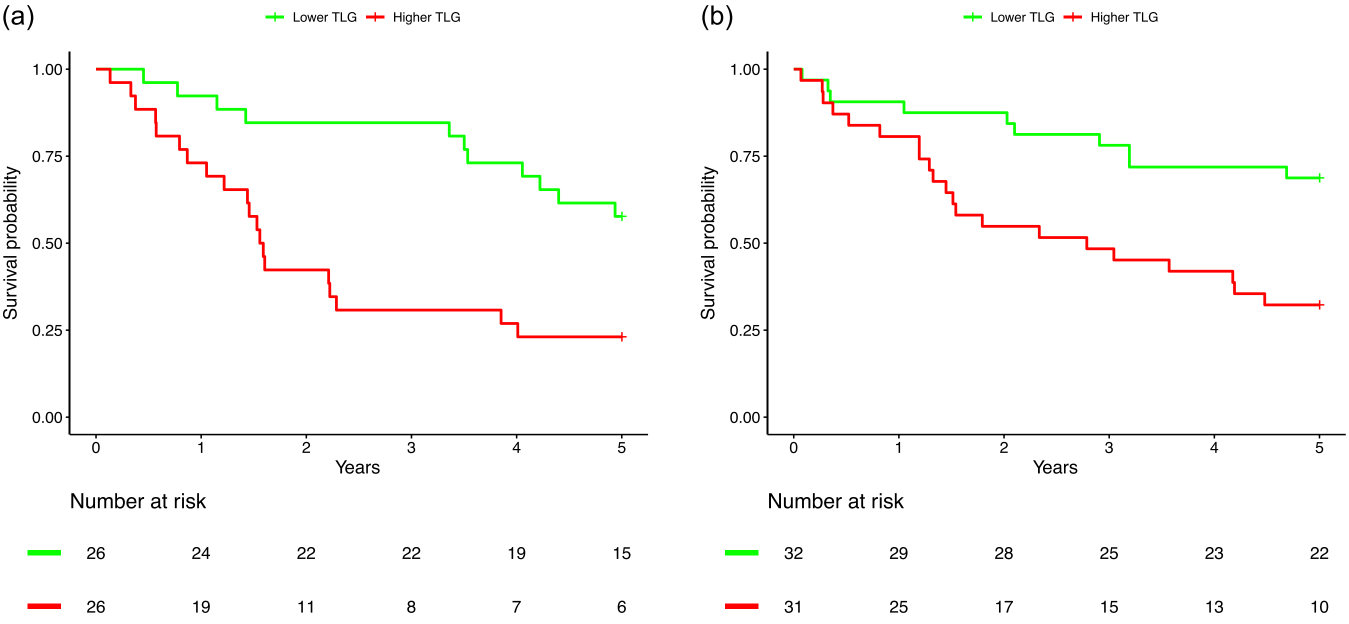

FIGURE 4 Kaplan–Meier curves showing survival probabilities for patients divided into two equally sized quantiles based on their PET-assisted reporting system (PARS) total lesion glycolysis (TLG). Patients with known lung cancer are shown to the left ($n = 52$) and suspected lung cancer to the right ($n = 63$).

manual TLGs were smaller in our study for lung tumours, which could be at least partly explained by lung tumours being easier to detect and segment than other soft-tissue tumours.

The road from a research prototype to a clinically accepted and approved AI tool includes several steps of validations and adjustments. Clinical validation of an AI tool should be based on test data with more advanced reference methods than those used in the present study and that of Sibilles et al. As an example, the process of qualifying the AI tool for the calculation of the Bone Scan Index (BSI) included several studies from the first presentation of the AI tool (Ulmert et al., 2012) to preanalytical (Anand, Morris, Kaboteh, Reza, et al., 2016), analytical (Anand, Morris, Kaboteh, Båth, et al., 2016), and clinical validations of the final version (Armstrong et al., 2018). The clinical validation was based on 721 patients from a phase III clinical trial and showed the association between BSI and OS.

The limitations of this study include the small number of patients and the use of PET/CT studies from an older scanner. The results from this scanner may not be representative of image quality obtained for more recent and advanced scanners. The retrospective design of the study allowed us to perform survival analysis but gave us limited access to clinical information such as other biomarkers, treatment and types of lung cancer.

In conclusion, this independent validation study of PARS shows that automatically calculated TLG contains prognostic information that is comparable to the manually measured TLG in patients with known or suspected lung cancer. The low sensitivity at both the lesion and patient levels makes the present version of PARS less useful to support clinical reading, reporting and staging.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Siemens Medical Solutions USA, Inc., for providing the research prototype to conduct this study. The authors are thankful to Vijay Shah, Mr. Ludovic Sibille, and Dr. med. Carl V. Gall for support with the protype. The Swedish state under the agreement between Swedish government and the county councils, the ALF-agreement (ALFGBG-720751).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. The images are not publicly available due to privacy or ethical restrictions.

ORCID

Pablo Borrelli http://orcid.org/0000-0002-1665-7088

Lars Edenbrandt http://orcid.org/0000-0002-0263-8820

REFERENCES

Anand, A., Morris, M.J., Kaboteh, R., Båth, L., Sadik, M., Gjertsson, P. et al. (2016) Analytic validation of the automated bone scan index as an imaging biomarker to standardize quantitative changes in bone scans of patients with metastatic prostate cancer. *Journal of Nuclear Medicine*, 57, 41–45.

Anand, A., Morris, M.J., Kaboteh, R., Reza, M., Trägårdh, E., Matsunaga, N. et al. (2016) A preanalytic validation study of automated bone scan index: effect on accuracy and reproducibility due to the procedural variabilities in bone scan image acquisition. *Journal of Nuclear Medicine*, 57, 1865–1871.

Armstrong, A.J., Anand, A., Edenbrandt, L., Bondesson, E., Bjartell, A. & Widmark, A. et al. (2018) Phase 3 assessment of the automated bone scan index as a prognostic imaging biomarker of overall survival in men with metastatic castration-resistant prostate cancer: a secondary analysis of a randomized clinical trial. *JAMA Oncology*, 4, 944–951.

Bluemke, D.A., Moy, L., Bredella, M.A., Ertl-Wagner, B.B., Fowler, K.J., Goh, V.J. et al. (2020) Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. *Radiology*, 294, 487–489.

Borrelli, P., Ly, J., Kaboteh, R., Ulén, J., Enqvist, O., Trägårdh, E. et al. (2021) AI-based detection of lung lesions in18F-FDG PET-CT from lung cancer patients. *EJNMMI Physics*, 8, 32.

Capobianco, N., Meignan, M., Cottereau, A.S., Vercellino, L., Sibille, L., Spottiswoode, B. et al. (2021) Deep-learning 18F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large b-cell lymphoma. *Journal of Nuclear Medicine*, 62, 30–36.

Groheux, D., Quere, G., Blanc, E., Lemarignier, C., Vercellino, L. & de Margerie-Mellon, C. et al. (2016) FDG PET-CT for solitary pulmonary nodule and lung cancer: literature review. *Diagnostic and Interventional Imaging*, 97, 1003–1017.

Heye, T., Merkle, E.M., Reiner, C.S., Davenport, M.S., Horvath, J.J., Feuerlein, S. et al. (2013) Reproducibility of dynamic contrast-enhanced MR imaging. part II. comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. *Radiology*, 266, 812–821.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H. & Aerts, H.J.W.L. (2018) Artificial intelligence in radiology. *Nature Reviews Cancer*, 18, 500–510.

Pinochet, P., Eude, F., Becker, S., Shah, V., Sibille, L., Toledano, N.M. et al. (2021) Evaluation of an automatic classification algorithm using convolutional neural networks in oncological positron emission tomography. *Frontiers in Medicine*, 8, 117.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P. et al. (2019) Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *Journal of the National Cancer Institute*, 111, 916–922.

Sibille, L., Seifert, R., Avramovic, N., Vehren, T., Spottiswoode, B., Zuehlsdorff, S. et al. (2020) 18 F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*, 294, 445–452.

Smith-Bindman, R., Kwan, M.L., Marlow, E.C., Theis, M.K., Bolch, W., Cheng, S.Y. et al. (2019) Trends in use of medical imaging in US health care systems and in Ontario, Canada, 2000-2016. *Journal of the American Medical Association*, 322, 843–856.

Smith-Bindman, R., Miglioretti, D.L. & Larson, E.B. (2008) Rising use of diagnostic medical imaging in a large integrated health system. *Health Affairs*, 27, 1491–1502.

Trägårdh, E., Borrelli, P., Kaboteh, R., Gillberg, T., Ulén, J., Enqvist, O. et al. (2020) RECOMIA-a cloud-based platform for artificial intelligence research in nuclear medicine and radiology. *EJNMMI Physics*, 7, 51.

Ulmert, D., Kaboteh, R., Fox, J.J., Savage, C., Evans, M.J., Lilja, H. et al. (2012) A novel automated platform for quantifying the extent of skeletal tumour involvement in prostate cancer patients using the bone scan index. *European Urology*, 62, 78–84.

Weber, M., Kersting, D., Umutlu, L., Schäfers, M., Rischpler, C., Fendler, W.P. et al. (2021) Just another "clever Hans"? Neural networks and FDG PET-CT to predict the outcome of patients with breast cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, 48, 3141–3150.

How to cite this article: Borrelli, P., Góngora, J. L. L., Kaboteh, R., Enqvist, O. & Edenbrandt, L. (2022) Automated classification of PET-CT lesions in lung cancer: an independent validation study. *Clinical Physiology and Functional Imaging*, 42, 327–332. https://doi.org/10.1111/cpf.12773