Journal of Nuclear Medicine, published on May 19, 2022 as doi:10.2967/jnumed.121.263598

Distinction of lymphoma from sarcoidosis at FDG PET/CT - evaluation of radiomic-feature guided machine learning versus human reader performance

Short running title: ML to distinguish sarcoidosis & lymphoma

Pierre Lovinfosse[1], Marta Ferreira[2] [*], Nadia Withofs[1], Alexandre Jadoul[1], Céline Derwael[1], Anne-Noelle Frix[3], Julien Guiot[3], Claire Bernard[1], Anh Nguyet Diep[4], Anne-Françoise Donneau[4], Marie Lejeune[5], Christophe Bonnet[5], Wim Vos[6], Patrick E. Meyer[7], Roland Hustinx[1]

1. Division of Nuclear Medicine and Oncological Imaging, CHU of Liège, Liège, Belgium.
2. GIGA-CRC in vivo Imaging, University of Liège, Liège, Belgium.
3. Department of Respiratory Medicine, CHU of Liège, Liège, Belgium.
4. Biostatistics Unit, Department of Public Health, University of Liège, Liège, Belgium.
5. Department of Hematology, CHU of Liège, Liège, Belgium
6. Radiomics SA, Liège, Belgium
7. Bioinformatics and Systems Biology Lab, University of Liège, Liège, Belgium.

[*] This author can be considered as first co-author

Corresponding author: Pierre Lovinfosse MD, PhD, Division of Nuclear Medicine and Oncological Imaging, University Hospital of Liège, 1, Avenue de l'Hôpital, 4000, Liège, Belgium. Telephone number: +3243667199, Fax number: +3243668533. E-mail: pierre.lovinfosse@chuliege.be

ABSTRACT

Sarcoidosis and lymphoma often share common features on ^{18}F-FDG PET/CT, such as intense hypermetabolic lesions of lymph nodes and multiple organs. We aimed at developing and validating radiomics signatures to differentiate sarcoidosis from Hodgkin (HL) and diffuse large B-cell (DLBCL) lymphoma. **Methods:** We retrospectively collected 420 patients (169 sarcoidosis, 140 HL and 111 DLBCL) who underwent a pretreatment ^{18}F-FDG PET/CT at the University Hospital of Liege. The studies were randomly distributed to 4 physicians who gave their diagnostic suggestion between the 3 diseases. Individual and pooled performances of physicians were then calculated. The inter-observer variability was evaluated using a sample of 34 studies interpreted by all physicians. Volumes of interest (VOI) were delineated over the lesions and the liver using MIM software, and 215 radiomic features were extracted using Radiomics toolbox. Models were developed combining clinical data (age, gender and weight) and radiomics (original and tumor-to-liver TLR radiomics), with 7 different feature selection approaches and 4 different machine learning (ML) classifiers, to differentiate sarcoidosis and lymphomas on both lesion-based and patient-based approaches. **Results:** For identifying lymphoma vs. sarcoidosis, physicians' pooled sensitivity, specificity, area under the curve (AUC) and accuracy were 0.99 (CI95%:0.97-1.00), 0.75 (CI95%: 0.68-0.81), 0.87 (CI95%: 0.84-0.90) and 89.3%, respectively, whereas for identifying HL in the tumor population, it was 0.58 (CI95%: 0.49-0.66), 0.82 (CI95%: 0.74-0.89), 0.70 (CI95%: 0.64-0.75) and 68.5%, respectively. A moderate agreement was found between observers for the diagnosis of lymphoma vs. sarcoidosis and HL vs. DLBCL with Fleiss kappa values of 0.66 (CI95%: 0.45-0.87) and 0.69 (CI95%: 0.45-0.93), respectively. The best ML models for identifying lymphoma vs. sarcoidosis showed AUC of 0.94 (CI95%: 0.93-0.95) and 0.85 (CI95%: 0.82-0.88) in lesion- and patient-based approaches, respectively, using TLR radiomics (+ age for the second). To differentiate HL and DLBCL, we obtained AUC of 0.95 (CI95%: 0.93-0.96) in lesion-based approach using TLR radiomics, and 0.86 (CI95%: 0.80-0.91) in patient-based using original radiomics and age. **Conclusion:** Characterization of sarcoidosis and lymphoma lesions is feasible using

ML and radiomics, with very good to excellent performances, equivalent or better than those of doctors who showed significant interobserver variability in their assessment.

Keywords: Radiomics, Machine Learning, Sarcoidosis, Lymphoma, ^{18}F-FDG PET/CT

INTRODUCTION

Sarcoidosis is a systemic inflammatory disease characterized by the development of granulomas that may involve lymph nodes and various organs. Hodgkin lymphoma (HL) and diffuse large B-cell lymphoma (DLBCL), the most frequent type of non-Hodgkin lymphoma, are also characterized by enlarged invaded lymph nodes but can also affect many organs. When 2-[^{18}F]fluoro-2-deoxy-D-glucose (^{18}F-FDG) PET/CT is performed at diagnosis, these diseases may present with a similar pattern, i.e. intense hypermetabolism in enlarged lymphadenopathies, in particular in the mediastinum. Involvement of many other nodal stations may also be observed, along with extranodal lesions, and the distribution of lesions thus helps imaging specialists in interpreting these PET/CT scans. Nonetheless the accuracy of the visual interpretation of ^{18}F-FDG PET/CT scans for differentiating sarcoidosis from lymphomas is imperfect (1). Semiquantitative measurements such as the maximum standardized uptake value (SUV$_{max}$) have not proven to be the answer either (2,3). Moreover, sarcoidosis can develop before (sarcoidosis-lymphoma syndrome) and after lymphoma, and immunotherapy-induced sarcoid-like reactions are increasingly observed (4-7). No matter the results of the imaging studies, pathological confirmation of the disease is mandatory in all cases prior to initiating treatment.

Histopathology of these entities is very different, suggesting that deep characteristics of the image might also be specific. Radiomics is a high-throughput approach allowing the extraction of large amounts of data from images and the characterization of the lesion phenotype (8,9). The development of artificial intelligence and machine learning (ML) combined with radiomics has gained popularity in different medical imaging tasks, including lesion identification and characterization. In lymphoma, some studies showed the potential of [^{18}F]FDG PET/CT radiomics to differentiate lymphoma from other types of cancers and to differentiate different types of lymphoma (10-14). To the best of our knowledge, no study has yet explored the use of ^{18}F-FDG PET/CT radiomics to characterize sarcoidosis lesions, except one for the diagnosis of cardiac involvement (15).

The primary objective of the present study was to develop and validate a radiomics signature to differentiate sarcoidosis, HL and DLBCL lesions. Furthermore, we compared the ML-driven diagnosis with the physicians' performance in categorizing the three diseases, taking into account the inter-observer variability.

MATERIALS AND METHODS

Patients

The study has been approved by the Ethics Committee of the University Hospital of Liège. The need for written informed consent was waived due to the retrospective and non-interventional design of the study. We retrospectively collected consecutive ^{18}F-FDG PET/CT scans performed at the University Hospital of Liège between 04/2010 and 02/2020 in patients with HL, DLBCL or sarcoidosis at initial diagnosis, prior to any treatment. The diagnosis was based on pathology in all lymphoma cases and in the large majority of cases of sarcoidosis. The diagnosis of the remaining sarcoidosis cases was based upon clinical evidence and follow-up. Exclusion criteria were radiotracer extravasation, artefacts in pathological areas, the absence of delineated volume of interest (VOI) after semi-automatic segmentation method (described below) and scans with missing relevant information in the DICOM (Digital imaging and communications in medicine) files. Basic clinical data (age, gender and weight) were collected from the information obtained routinely on the day of the PET/CT. Figure 1 shows the flowchart of the study.

Imaging

^{18}F-FDG PET/CT scans were acquired using 2 cross-calibrated PET/CT systems: a GEMINI TF Big Bore and a GEMINI TF 16 (Philips Medical Systems, Cleveland, OH, USA); 66 minutes in average (range: 58-92) after intravenous injection of ^{18}F-FDG (mean injected activity: 245 MBq, depending linearly on patient's weight). Patients fasted for at least 6h before the injection and the median glycemia was 92 mg/dl (range: 59-195). A low-dose CT (5 mm slice thickness; tube voltage: 120 kV and tube current–time product: 50 to 80 mAs depending on the patient's weight) was performed without injection of intravenous contrast agent, followed by a PET emission scan of 90 seconds per bed position (50% of overlapping) extending from the upper thighs to the skull base. All images were acquired and reconstructed according to the European Association of Nuclear Medicine Research Limited guidelines for both PET/CT systems. Images were reconstructed with standard 4x4x4 mm^3 voxels (slice thickness 4 mm) using iterative list mode time-of-flight algorithms (BLOB-OS-TF) and corrections for attenuation, dead-time, random and scatter events were applied, without post-reconstruction smoothing.

Lesion segmentation and clinical diagnosis

The entire cohort of anonymized patients was randomly distributed into 4 groups (A to D) and attributed to 4 nuclear medicine physicians, unaware of any clinical information or diagnosis, with an experience of 6 years (observer A), 3 years (observer B), 15 years (observer C) and 10 years (observer D). In a first step, based on the visual interpretation of the PET/CT, the physicians attributed a diagnosis to each patient of their cohort. For that purpose, they first assigned either sarcoidosis or cancer and, if the latter was selected, they chose HL or DLBCL. This evaluation was solely based on the experience of each physician. No reading guidelines, visual or semi-quantitative interpretation criteria were provided to the readers within the framework of the study. For each answer, the physicians indicated their level of confidence: 0=possible, 1=probable and 2=certain.

In the next step, every physician segmented PET VOIs of his/her assigned patients population using MIM software v7.0.5 (MIM Software Inc, Cleveland, OH) with the following procedure: 1) automatic selection of all the regions using an absolute threshold $SUV_{max} \geq 3$ within a rectangular VOI manually drawn on the whole-body, 2) automatic exclusion of VOIs smaller than 2 cc, 3) manual exclusion of all the physiological VOIs (brain, heart, kidneys, …), 4) manual modification of some pathological VOIs, i.e. removing physiological activity in continuity with the pathological VOI, never enlarging the VOI. In the absence of literature references for this combination of diseases, especially considering sarcoidosis, the thresholds of SUV_{max} and volume were decided after tests were performed on a sample of images with the aim of including as many lesions as possible while limiting the need for manual modifications. A VOI of 20 ml was also drawn in the healthy liver.

Radiomics extraction and models elaboration

Two hundred and fifteen features were extracted from the segmented PET volumes using the Radiomics research toolbox (Radiomics SA, Liège, Belgium), coded with Matlab and which is aligned with the imaging biomarkers standardization initiative (IBSI) with however some additional features (list of all features in supplemental material). We also studied the ratio of the features values calculated in the tumor and in the liver (TLR), except for the shape features. For the calculation of the texture matrix-based features, the intensities were discretized using two different methods according to IBSI recommendations: fixed bin number, using 32 and 64 bins, and fixed bin width with 4 different widths of 0.05, 0.1, 0.2 and 0.5 SUV.

Since each patient may have more than one lesion, two radiomics approaches were tested. In a first approach, each lesion was considered as one observation (‘lesion-based approach’) and the goal was to classify each lesion as, firstly, belonging to the sarcoidosis or lymphoma class and secondly, as belonging to the HL or DLBCL class. In the second approach (‘patient-based approach’), radiomic features of each lesion and for each patient were merged using 1) their minimum value 2) their maximum value 3) their mean value 4) their median value, and clinical data (age, gender and weight) were added to the radiomic features. Here, the aim was to classify each patient into the sarcoidosis, HL or DLBCL groups.

We also evaluated whether combining different feature selection (FS) approaches and ML classifiers would allow for a radiomics signature 1) to differentiate sarcoidosis from lymphoma patients and 2) to differentiate HL from DLBCL. For that purpose, we tested a different set of models, which differ in i) the features type, i.e., original radiomics (OR) or TLR radiomics, ii) the FS and ML classifier method, and iii) the effect of adding clinical data before FS. Seven different FS methods were tested: 1-Accuracy decrease obtained from the embedded FS of the random forest (RF) classifier; 2-Gini impurity decrease obtained from the embedded FS of the RF classifier; 3- Forward FS using Minimum Redundancy Maximum Relevance (MRMR) method with Pearson correlation; 4- Backward FS using MRMR with Pearson correlation; 5- Forward FS using MRMR with Spearman correlation; 6-Backward FS using MRMR with Spearman correlation; 7- Forward MRMR based on the mutual information. We also used 4 ML classifiers: RF, support vector machine with radial kernel, naive Bayes (NB) and a logistic regression(16). The dataset was stratified with the same percentage of classes, avoiding unbalanced data, and randomly divided into training and test sets (80% and 20%, respectively). We tested different models that differ in the FS, ML and intensity discretization method as well as the number of features, which was between 2 to 20 with intervals of two. We used 5-fold cross validation in our training data, and we chose the best radiomic signature according to the best mean 5-fold cross validation area under the precision-recall curve (AUCpr). For each classifier, the default hyperparameters values were used in their respective R packages. Finally, for each of the different models with distinct selected features, all training data were bootstrapped, in order to derive the corresponding 95% confidence intervals for each performance metric and tested on the independent test set. The number of bootstrap repetitions was set to at least 1000 repetitions. Since images came from only two scanners (same manufacturer and model, same acquisition protocols), which were also cross-calibrated, we did not consider necessary to perform data harmonization. As recently suggested by Buvat and Orlhac (17), we performed a T.R.U.E checklist to assess the potential impact of our findings.

Statistical analysis

The homogeneity in terms of age and weight across the A, B, C, and D populations was assessed by Kruskal-Wallis non-parametric test, whereas chi-square test association was performed for gender and the final diagnosis (gold standard). Additionally, the homogeneity in terms of age, gender and weight across cancer and sarcoidosis patients as well as HL vs. DLBCL patients was also evaluated using chi-square test and Kruskal-Wallis test. Statistical significance was assigned for *p*-value lower or equal to 0.05.

The diagnostic performances of all observers pooled together and each individual observer against the gold-standard was calculated using sensitivity (Se), specificity (Sp), positive predictive value (PPV) and negative predictive value (NPV). In addition, the diagnostic performance was evaluated by calculating the percentage of agreement (or accuracy) and the AUC. To calculate the predicted probabilities, we fitted a logistic regression model with the observer's classification as the predictor. We additionally bootstrapped the data to measure the AUC 95% confidence intervals. The confidence intervals for the Se, Sp, PPV, and NPV were calculated using exact binomial confidence limits.

To test the variability between observers, we applied the confidence interval approach in sample size estimation for inter-observer agreement with binary outcomes (18). Due to a lack of literature on the established agreement, to calculate the sample size in the case of 4 observers, we assigned an expected kappa of 0.70, indicating moderate agreement (19), a lower bound of 0.50 with unknown upper bound, and a significance level of 0.05. With a prevalence of 0.3 for the sarcoidosis vs. cancer and 0.6 for the HL vs. DLBCL classifications, a sample size of 27 and 23 was required, respectively. Based on this estimation, we randomly selected a subgroup of 34 patients, who were subsequently analyzed by all 4 observers, in order to evaluate the inter-observer variability. Due to the misclassification for the sarcoidosis vs. cancer, only 21 patients remained in the evaluation of inter-observer variability in the analysis of HL vs. DLBCL. Fleiss's kappa was employed to investigate the overall agreement between 4 observers in the classification (for sarcoidosis/cancer, and for HL/DLBCL) and intraclass correlation coefficients (ICCs) for degree of certainty. Finally, Hotelling's T-squared (T^2) test was used to test the difference in agreement between pairs of observers.

For radiomics, we evaluated performances of the models described above using AUC, AUCpr, Se, Sp, PPV and NPV with 0.5 probabilities threshold on the test set, for lesion-based and patient-based approaches. Statistical and ML analyses were performed for the two clinical tasks (sarcoidosis/cancer and HL/DLBCL) using R software V4.0.3.

RESULTS

A total of 448 patients meeting the study inclusion criteria were initially identified. After applying the exclusion criteria, 420 patients (mean age 49±18y; 241/179 men/women) remained in the study (Figure 1). According to the gold-standard, 169 patients had sarcoidosis (40.2%), 140 HL (33.3%) and 111 DLBCL (26.4%). Ann Arbor stages for HL and DLBCL were 10 I, 1 IE, 64 II, 1 IIE, 19 III, 20 IIIS, 25 IV and 10 I, 27 II, 12 III, 10 IIIS, 52 IV, respectively; and 81 patients with sarcoidosis had extra-thoracic lesions. Table 1 presents patient's characteristics and gold-standard for the entire population and the 4 physicians' subsamples. The 4 groups were balanced except for observer D, who had significantly more sarcoidosis patients and fewer DLBCL. A significant difference in patients' age was observed across the 4 subsamples (*p*=0.008), with patients in group D significantly younger than those in groups A and B; which could be explained by the difference in the distribution of diseases. There was no significant difference for weight and gender across the 4 subsamples.

Individual and pooled observer's performances as compared to the gold-standard

For identifying lymphomas (HL & DLBCL) in the entire population (n=420), the Se and Sp were 0.99 (0.97-1.00) and 0.75 (0.68-0.81), respectively. AUC-ROC was 0.87 (0.84-0.90) and accuracy was 0.893 (0.86-0.92). Similarly, Cohen's kappa (κ)=0.78 (0.72-0.84) revealed a substantial agreement with the gold standard. Taking the certainty level into account, a significant higher agreement κ=0.86 (0.79-0.92) was found for certainty level 2 compared to a κ=0.41 (0.23-0.58) for level 1 (*p*<0.001).

Overall and individual observers' performances for the diagnosis of cancer versus sarcoidosis for their sub-sample populations are listed in Table 2. All observers had an excellent Se (0.97 to 1.00) but a lower and more variable Sp (0.58 to 0.81). AUC and accuracy ranged from 0.79 to 0.90 and from 0.85 to 0.92, respectively. Regarding the confidence levels, observers A, B, C and D chose the level 2 in 81%, 80%, 80% and 65% of cases; level 1 in 15%, 19%, 12% and 27% of cases; and level 0 in 4%, 1 %, 8% and 8% of cases, respectively.

For identifying HL in the cancer population (n=248, after removing 3 patients mistakenly categorized with sarcoidosis), the Se and Sp were 0.58 (0.49-0.66) and 0.82 (0.74-0.89) respectively. AUC-ROC was 0.70 (0.64-0.75) and accuracy was 0.69 (0.63-0.74). Cohen's kappa κ=0.40 (0.29-0.51) indicated only a fair agreement with the gold standard. When certainty level was at 2, a significantly higher κ=0.51 (0.41-0.67) was obtained compared to a κ=0.20 (0.14-0.39) at level 1 of certainty (*p*=0.003).

Overall and individual observers' performances for the diagnosis of HL vs. DLBCL for their sub-sample populations are listed in Table 3. The Se ranged from 0.39 to 0.77 and Sp from 0.77 to 0.85. AUC and accuracy ranged from 0.60 to 0.81 and from 0.59 and 0.82, respectively. Regarding the confidence levels, observers A, B, C and D selected the level 2 in 54%, 46%, 61% and 45% of cases; level 1 in 38%, 49%, 33% and 43% of cases and level 0 in 8%, 5%, 6% and 12% of cases, respectively. Representative examples of PET studies are shown in Figures 2 and 3.

Inter-observer agreement

In the sample of 34 patients, a Fleiss' kappa value of 0.66 (0.45-0.87) indicated that the four observers were in moderate agreement with one another in the diagnosis of cancer vs. sarcoidosis. Regarding the certainty levels, an ICC=0.353 (0.181-0.547) showed poor agreement among the observers. At the individual level, the agreement with the gold standard was highly variable, as κ ranged from 0.45 to 0.93. Hotelling's T^2 test showed that the agreement with the gold standard differed significantly between the 2 extreme values, i.e. observers B and D ($T^2 = 8.70$, *p*=0.006).

For the diagnosis of HL versus DLBCL, in the population of 21 patients diagnosed with cancer evaluated by all four observers, the Fleiss' kappa value of 0.69 (0.45-0.93) indicated a moderate agreement between observers. Regarding the certainty levels, an ICC of 0.075 (0.076-0.316) showed poor agreement among the observers. At the individual level, only observer A displayed a substantial agreement with the gold-standard (κ=0.70; 0.38-1.01), whereas the three other observers showed poor agreement with κ ranging from 0.07 to 0.27. Hotelling's T^2 test showed that observer A outperformed the other three observers in terms of agreement with the gold-standard and the most significant difference was between observers A and B (T^2=9.60, *p*=0.006). There was no significant difference of agreement between observers B, C and D. Supplemental tables 1-2 show all the individual kappa and Hotelling's T^2 values of the inter-observer agreement analysis, for the two tasks.

Performance of radiomics models compared to the gold-standard

In the whole cohort, 2816 VOIs were segmented, including 1028 (36.5%) for sarcoidosis, 836 for HL (29.7%) and 952 (33.8%) for DLBCL (mean number of VOIs by patient: 42.1 for sarcoidosis, 44.7 for HL and 75.8 for DLBCL). One patient with sarcoidosis was excluded from the radiomics analyses (n=419) because of diffuse liver pathological infiltration that did not allow the delineation of the hepatic background VOI. The results of the best models compared to the physicians' performances are summarized in Figures 4 and 5.

A RF classifier, where features were selected with the embedded RF features selection using the accuracy decrease as a criterion, yielded the best performances to differentiate cancer from sarcoidosis following a 'lesion-based' approach. This model included 4 TLR radiomics features discretized with FBW with width of 0.05 SUV: two first-order gray-level statistics features (Stats_min; Stats_p10), one intensity volume histogram feature (IVH_AIRV_90) and one textural feature (GLCM_infoCorr2). This model showed Se of 0.92 (0.89-0.94), Sp of 0.80 (0.75-0.84), PPV of 0.88 (0.86-0.91) and NPV of 0.85 (0.81-0.89). For the test set, performances were excellent with AUC and AUCpr of 0.94 (0.93-0.95) and 0.96 (0.95-0.97), respectively, and it performed significantly better than the best model with original radiomics (AUC 0.68 and AUCpr 0.78). The best "patient-based" radiomics models included TLR radiomics (intensity volume histogram, shape and texture features), merged using their minimum values, and age of patients but showed poorer results than differentiation by lesion, with AUC and AUCpr of 0.85 (0.82-0.88) and 0.88 (0.84-0.92), respectively. For a decisional threshold of 0.5, it showed Se of 0.84 (0.78-0.90), Sp of 0.67 (0.56-0.76), PPV of 0.79 (0.74-0.84) and NPV of 0.74 (0.67-0.83), respectively. Supplemental tables 3-6 shows the selected features and results of the best original and TLR radiomics models for lesion-based and patient-based analysis.

To differentiate HL and DLBCL, the 'lesion-based' radiomics model with the best performances used RF classifier (Gini impurity decrease) and was composed of 2 TLR radiomics features discretized with FBW with width of 0.05 SUV: one first-order gray-level statistics features (Stats_min) and one textural feature (GLCM_infoCorr2). It showed Se, Sp, PPV and NPV of 0.89 (0.85-0.92), 0.88 (0.84-0.92), 0.87 (0.83-0.90) and 0.90 (0.87-0.92), respectively. For the test set, performances were excellent with AUC and AUCpr of 0.95 (0.93-0.96) and 0.95 (0.92-0.96), respectively, close to those of the validation set (AUC and AUCpr of 0.97, both) and significantly better than the best model with original radiomics (AUC 0.67 and AUCpr 0.62). The best patient-based radiomics models used a NB classifier and a forward MRMR with Pearson correlation for feature selection. The model included original radiomic features merged with their maximal values and discretized with FBW with width of 0.5 SUV (first-order, intensity volume histogram, and textural features: IH-entropy, IVH_AIRV_70, GLCM_infoCorr1, NGLDM_SM, NGLDM_DNN) and patient's age. It showed very good performances with AUC and AUCpr of 0.86 (0.80-0.91) and 0.87 (0.78-0.91), respectively. For a decisional threshold of 0.5, this model showed Se of 0.79 (0.71-0.86), Sp of 0.85 (0.73-0.86), PPV of 0.87 (0.79-0.89) and NPV of 0.76 (0.70-0.83), respectively.

DISCUSSION

In cancer imaging, ^{18}F-FDG PET/CT takes advantage of a high sensitivity but the specificity is intrinsically limited by significant uptake by various inflammatory and infectious lesions. Obviously ^{18}F-FDG uptake alone cannot reliably identify the pathology of the tumor. In this study, we developed radiomics signatures to characterize lesions with highly increased ^{18}F-FDG uptake, as a proof of concept of machine learning to differentiate inflammation from cancer, and to differentiate 2 cancer types. At the lesion level, we found highly accurate signatures with an AUC of 0.94 for the first task, and 0.95 for the second one. At the patient level, we created models with very good performances to differentiate cancer vs. sarcoidosis (AUC 0.85) and HL vs. DLBCL (AUC 0.86), which were respectively equivalent and significantly better than human performances. All physicians showed an excellent sensitivity (0.97-1.00) to identify patients with cancer and a good but lower specificity (0.75). Overall, the global performances were good with an AUC of 0.87. However, there was only a moderate agreement among the observers, especially due to poorer performance of the youngest observer (resident in training with 3 years of experience). Furthermore, the observers greatly varied in their level of certainty when deciding whether a PET/CT scan results was cancer or sarcoidosis. Interestingly enough, this was with a significant correlation with performance, i.e. higher confidence was associated with better performances. To differentiate HL and DLBCL, the overall performance of the physicians deteriorated with an AUC of 0.70, which was related to a moderate sensitivity. Again, large variability was observed among the observers with one of them performed significantly better than the other three. However, the difference was unrelated to the experience, whereas a significant correlation was observed with the degree of certainty. Observer D had a different sample of diseases compared to the others. Yet, the fact that he was not aware of this information and his performances in his subsample and in the inter-observer variability analysis were reassuring factors as to this possible confounding effect on the obtained results.

The findings confirmed that radiomic analysis of the metabolic signal could effectively distinguish inflammatory and neoplastic lesions (20-22) but also different types of cancer (10,12,23-26). Regarding lymphomas, in a population of 25 patients, Lartizien et al. used ^{18}F-FDG PET/CT radiomics and support vector machine classifier to distinguish aggressive lymphoma lesions (B-cell lymphoma and HL) from non-lymphomatous uptake sites (brown fat, inflammation, infection, physiologic thymic uptake, …) with an AUC of 0.91 (27). Lippi et al. related good performance of ML to discriminate different types of lymphomas from each other, especially HL, but in a small population of patients (11). Recently, de Jesus et al. showed very promising results to differentiate follicular lymphoma and DLBCL using radiomics and ML classifier in a population of 120 patients, which could have important clinical use when monitoring patients for aggressive transformation (14). Their best performing model showed an AUC of 0.86 significantly higher than the performance of the SUV_{max}-based model (AUC 0.79). In addition to the significant difference of population size and different types of lymphoma, certain methodological differences should be highlighted with our work, including the type of ML classifier based on per-lesion only, the segmentation method and choice of analyzed lesions, the absence of comparison with human performance and the use of radiomics of PET and CT simultaneously. Beyond the proof of concept, our results may have clinical implications. Indeed, the high sensitivity of the model could avoid an invasive biopsy procedure in patients with sarcoidosis, providing these excellent results be confirmed in a large and independent external population.

Machine learning algorithms performances depend on several factors, including, but not limited to: data size, randomness during learning or pre-processing steps (28). Given this, we tested a different set of models, which differs in i) the feature types, i.e. original radiomics or TLR radiomics, ii) the FS strategy and number of features, iii) the intensity discretization scheme. We have shown in previous studies that using the ratio of the tumor features with the liver as reference organ improves the predictive performance in cervical cancer as well as the robustness across centers (16). The improvement in models' performance can be caused by a normalizing effect of the SUVs on each patient. In the present study, the TLR models systematically outperformed the OR models in the 'lesion-based approach' but not when considering the 'patient-based' approach. Despite of this fact, the performances of the models when using TLR features were close to the ones using the original features, showing the high potential of ratio-based features in terms of applicability in different centers.

Even though the present study followed the IBSI standardization initiative guidelines and scored 56% according to the radiomics quality score (29), it has several limitations, including its retrospective and monocentric design, with the need for an external validation within an independent population. It is possible that performances of physicians were underestimated in comparison to clinical routine due to the complete absence of clinical data. Moreover, physicians were nuclear medicine specialists without specific training in radiology, which could potentially influence performance. Conversely, the performance of radiomics and ML could possibly be improved by integrating more clinical (sweats, weight loss, etc…) and biological data, the localization of lesions (11), the CT or MRI radiomics (14,24,30)and by using a deep learning approach (31). In our study, some VOIs were manually adapted in case of physiological activity overflowing on a pathological VOI. However, those were rare occurrences, which was less likely to result in biases in the results. Also, we excluded from the study the patients without any VOI generated by the automated segmentation process. Given that these represented only a small part of the population (n=12/448 patients; 2.5 %), it was unlikely that they would have affected the results. Finally, to show validity, reproducibility, usefulness and explainability of our results, we add a T.R.U.E checklist in supplemental material.

CONCLUSION

The characterization of sarcoidosis and lymphoma lesions is feasible using ML and radiomics, inherent in their very good to excellent performances, proving to be equivalent or better than those of doctors who showed significant interobserver variability in their assessment.

DISCLOSURES

There were no personal grants, consulting fees, or honoraria involved in the present paper. No potential conflicts of interest relevant to this article exist. This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 766276. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Dr. Adeline Deward from the GIGA scientific illustration service for designing the graphical abstract.

KEY POINTS

1) Question: Are specialists in medical imaging able to differentiate sarcoidosis and lymphomas based on visual analysis of ^{18}F-FDG PET/CT and can machine learning models using radiomics help them in this task?
2) Pertinent findings: Physicians characterize these diseases with variable performance, from moderate to very good. Machine learning and radiomics models achieve similar and better performances, in a more reproducible way.
3) Implications for patient care: Machine learning and radiomics models are able to differentiate sarcoidosis and lymphoma, making it possible to consider, after external validation, their use in order to avoid unnecessary biopsies in patients with high suspicion of sarcoidosis.

REFERENCES

1. Li YJ, Zhang Y, Gao S, Bai RJ. Cervical and axillary lymph node sarcoidosis misdiagnosed as lymphoma on F-18 FDG PET-CT. *Clin Nucl Med.* 2007;32:262-264.

2. Koo HJ, Kim MY, Shin SY, et al. Evaluation of mediastinal lymph nodes in sarcoidosis, sarcoid reaction, and malignant lymph nodes using CT and FDG-PET/CT. *Medicine (Baltimore).* 2015;94:e1095.

3. Yu C, Xia X, Qin C, Sun X, Zhang Y, Lan X. Is SUVmax helpful in the differential diagnosis of enlarged mediastinal lymph nodes? A pilot study. *Contrast Media Mol Imaging.* 2018;2018:3417190.

4. Brady B, Kamel D, Kiely J, Hennessy B. Dual diagnosis of sarcoidosis and lymphoma. *Ir J Med Sci.* 2013;182:283-286.

5. Sanan P, Lu Y. Multiorgan involvement of chemotherapy-induced sarcoidosis mimicking progression of lymphoma on FDG PET/CT. *Clin Nucl Med.* 2017;42:702-703.

6. Bando-Delaunay A, Luporsi M, Huchet V, Cassou-Mounat T, Jehanno N. A case of sarcoidosis after lymphoma. *Clin Nucl Med.* 2019;44:646-647.

7. Cayci Z, Ozturk K, Ustun C, et al. Sarcoid-like histiocytic proliferations in patients with lymphoma can be FDG-avid concerning for refractory or recurrent disease. *Clin Lymphoma Myeloma Leuk.* 2019;19:e597-e601.

8. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012;48:441-446.

9. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278:563-577.

10. Kong Z, Jiang C, Zhu R, et al. (18)F-FDG-PET-based radiomics features to distinguish primary central nervous system lymphoma from glioblastoma. *Neuroimage Clin.* 2019;23:101912.

11. Lippi M, Gianotti S, Fama A, et al. Texture analysis and multiple-instance learning for the classification of malignant lymphomas. *Comput Methods Programs Biomed.* 2020;185:105153.

12. Ou X, Zhang J, Wang J, et al. Radiomics based on (18) F-FDG PET/CT could differentiate breast carcinoma from breast lymphoma using machine-learning approach: A preliminary study. *Cancer Med.* 2020;9:496-506.

13. Zhu S, Xu H, Shen C, et al. Differential diagnostic ability of 18F-FDG PET/CT radiomics features between renal cell carcinoma and renal lymphoma. *Q J Nucl Med Mol Imaging.* 2021;65:72-78.

14. de Jesus FM, Yin Y, Mantzorou-Kyriaki E, et al. Machine learning in the differentiation of follicular lymphoma from diffuse large B-cell lymphoma with radiomic [(18)F]FDG PET/CT features. *Eur J Nucl Med Mol Imaging.* 2021;In press.

15. Manabe O, Ohira H, Hirata K, et al. Use of (18)F-FDG PET/CT texture analysis to diagnose cardiac sarcoidosis. *Eur J Nucl Med Mol Imaging.* 2019;46:1240-1247.

16. Ferreira M, Lovinfosse P, Hermesse J, et al. [(18)F]FDG PET radiomics to predict disease-free survival in cervical cancer: a multi-scanner/center study with external validation. *Eur J Nucl Med Mol Imaging.* 2021;48:3432-3443.

17. Buvat I, Orlhac F. The T.R.U.E. checklist for identifying impactful artificial intelligence-based findings in nuclear medicine: is it true? is it reproducible? is it useful? is it explainable? *J Nucl Med.* 2021;62:752-754.

18. Rotondi MA, Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. *J Clin Epidemiol.* 2012;65:778-784.

19. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012;22:276-282.

20. Du D, Gu J, Chen X, et al. Integration of PET/CT radiomics and semantic features for differentiation between active pulmonary tuberculosis and lung cancer. *Mol Imaging Biol.* 2021;23:287-298.

21. Hu Y, Zhao X, Zhang J, Han J, Dai M. Value of (18)F-FDG PET/CT radiomic features to distinguish solitary lung adenocarcinoma from tuberculosis. *Eur J Nucl Med Mol Imaging.* 2021;48:231-240.

22. Liu Z, Li M, Zuo C, et al. Radiomics model of dual-time 2-[(18)F]FDG PET/CT imaging to distinguish between pancreatic ductal adenocarcinoma and autoimmune pancreatitis. *Eur Radiol.* 2021;31:6983-6991.

23. Kirienko M, Cozzi L, Rossi A, et al. Ability of FDG PET and CT radiomics features to differentiate between primary and metastatic lung lesions. *Eur J Nucl Med Mol Imaging.* 2018.

24. Sibille L, Seifert R, Avramovic N, et al. (18)F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks. *Radiology.* 2020;294:445-452.

25. Ren C, Zhang J, Qi M, et al. Machine learning based on clinico-biological features integrated (18)F-FDG PET/CT radiomics for distinguishing squamous cell carcinoma from adenocarcinoma of lung. *Eur J Nucl Med Mol Imaging.* 2021;48:1538-1549.

26. Zhou Y, Ma XL, Zhang T, Wang J, Zhang T, Tian R. Use of radiomics based on (18)F-FDG PET/CT and machine learning methods to aid clinical decision-making in the classification of solitary pulmonary lesions: an innovative approach. *Eur J Nucl Med Mol Imaging.* 2021;48:2904-2913.

27. Lartizien C, Rogez M, Niaf E, Ricard F. Computer-aided staging of lymphoma patients with FDG PET/CT imaging based on textural information. *IEEE J Biomed Health Inform.* 2014;18:946-955.

28. Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci.* 2021;2:160.

29. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol.* 2017;14:749-762.

30. Santos FS, Verma N, Marchiori E, et al. MRI-based differentiation between lymphoma and sarcoidosis in mediastinal lymph nodes. *J Bras Pneumol.* 2021;47:e20200055.

31. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med.* 2021;4:65.

Tables and Figures

Table 1. Patients' characteristics for the entire population (n=420) and the 4 physicians subsamples

	Overall	**Observer A**	**Observer B**	**Observer C**	**Observer D**
Age (y),median (Q1-Q3)	49 (35-61)	52 (36-67)	52 (37-61)	49 (39-60)	44 (29-55)
Weight (Kg),median (Q1-Q3)	75 (63-86)	74 (62-84)	75 (66-85)	72 (62-85)	77 (63-89)
Gender: female-male	179-241	47-62	41-61	45-55	46-64
Diagnosis					
Sarcoidosis	169 (40.2%)	36 (33%)	36 (35.5%)	34 (34%)	63 (57%)
Hodgkin lymphoma	140 (33.3%)	32 (29%)	36 (35.5%)	35 (35%)	37 (34%)
Diffuse large B-cell lymphoma	111 (26.5%)	41 (38%)	29 (29%)	31 (31%)	10 (9%)

Table 2. Overall and individual observers' performances for the diagnosis of sarcoidosis versus lymphoma. Between brackets: 95% confidence intervals.

	Overall	**Observer A**	**Observer B**	**Observer C**	**Observer D**
Proposed diagnosis: Sarcoidosis-Cancer	130-290	31-78	21-80	28-72	50-60
Correct classification	375/420: 89.3% (86.3-92.2%)	100/109: 91.7% (86.6-96.9%)	86/101: 85.1% (78.2-92.1%)	92/100: 92% (86.7-97.3%)	97/110: 88.2% (82.3-94.2%)
Correct sarcoidosis classification	133/169: 78.7% (72.5-84.9%)	29/36: 80.6% (67.6-93.5%)	21/36: 58.3% (42.2-74.4%)	27/34: 79.4% (65.8-93%)	56/63: 88.9% (81.1-96.7%)
Correct cancer classification	248/251: 98.8% (97.5-100%)	71/73: 97.3% (93.5-100%)	65/65: 100%	65/66: 98.5% (95.5-100%)	47/47: 100%
Sensitivity	0.99 (0.97-1.00)	0.97 (0.90-1.00)	1.00 (0.94-1.00)	0.98 (0.92-1.00)	1.00 (0.92-1.00)
Specificity	0.75 (0.68-0.81)	0.81 (0.64-0.92)	0.58 (0.41-0.74)	0.79 (0.62-0.91)	0.79 (0.67-0.89)
Positive predictive value	0.86 (0.81-0.89)	0.91 (0.81-0.96)	0.81 (0.71-0.89)	0.90 (0.82-1.00)	0.78 (0.66-0.88)
Negative predictive value	0.98 (0.93-1.00)	0.94 (0.79-0.99)	1.00 (0.84-1.00)	0.96 (0.82-1.00)	1.00 (0.93-1.00)
AUC-ROC	0.87 (0.84-0.90)	0.89 (0.82-0.96)	0.79 (0.71-0.87)	0.89 (0.82-0.96)	0.90 (0.85-0.95)

Table 3.Overall and individual observers' performances for the diagnosis of Hodgkin (HL) versus Diffuse Large B-Cell (DLBCL) lymphomas. Between brackets: 95% confidence intervals.

	Overall	**Observer A**	**Observer B**	**Observer C**	**Observer D**
Proposed diagnosis: HL-DLBCL	110-180	33-45	22-58	27-45	28-32
Correct HL classification	80/140: 57.1% (49.0-65.3%)	23/32: 71.9% (56.3-87.5%)	14/36: 38.9% (23.0-54.8%)	20/35: 57.1% (40.8-73.5%)	27/37: 73.0% (58.7-87.3%)
Correct DLBCL classification	91/111: 82% (74.8-89.1%)	35/41: 85.4% (74.5-96.2%)	24/29: 82.8% (69.0-96.5%)	24/31: 77.4% (62.7-92.1%)	8/10: 80% (55.2-100%)
When observer said cancer and gold-standard was cancer:					
- correct HL classification	79/137: 57.7% (49.4-65.9%)	23/30: 76.7% (61.5-91.8%)	14/36: 38.9% (23.0-54.8%)	20/34: 58.8% (42.3-75.4%)	22/37: 59.5% (43.6-75.3%)
- correct DLBCL classification	91/111: 82% (74.8-89.1%)	35/41: 85.4% (74.6-96.2%)	24/29: 82.8% (69.0-96.5%)	24/31: 77.4% (62.7-91.1%)	8/10: 80% (55.2-100%)
Sensitivity	0.58 (0.49-0.66)	0.77 (0.58-0.90)	0.39 (0.23-0.57)	0.59 (0.41-0.75)	0.59 (0.42-0.75)
Specificity	0.82 (0.74-0.89)	0.85 (0.71-0.94)	0.83 (0.64-0.94)	0.77 (0.59-0.90)	0.80 (0.44-0.97)
Positive predictive value	0.80 (0.71-0.87)	0.79 (0.60-0.92)	0.74 (0.49-0.91)	0.74 (0.54-0.89)	0.92 (0.73-0.99)
Negative predictive value	0.61 (0.53-0.69)	0.83 (0.69-0.93)	0.52 (0.37-0.67)	0.63 (0.46-0.78)	0.35 (0.16-0.57)
Accuracy	170/248: 68.5% (62.7-74.3%)	58/71: 81.7% (72.7-90.7%)	38/65: 58.5% (46.5-70.5%)	44/65: 67.7% (56.3-79.1%)	30/47: 63.8% (50.1-77.5%)

AUC-ROC	0.70 (0.64-0.75)	0.81 (0.72-0.91)	0.60 (0.50-0.72)	0.68 (0.57-0.79)	0.70 (0.54-0.85)

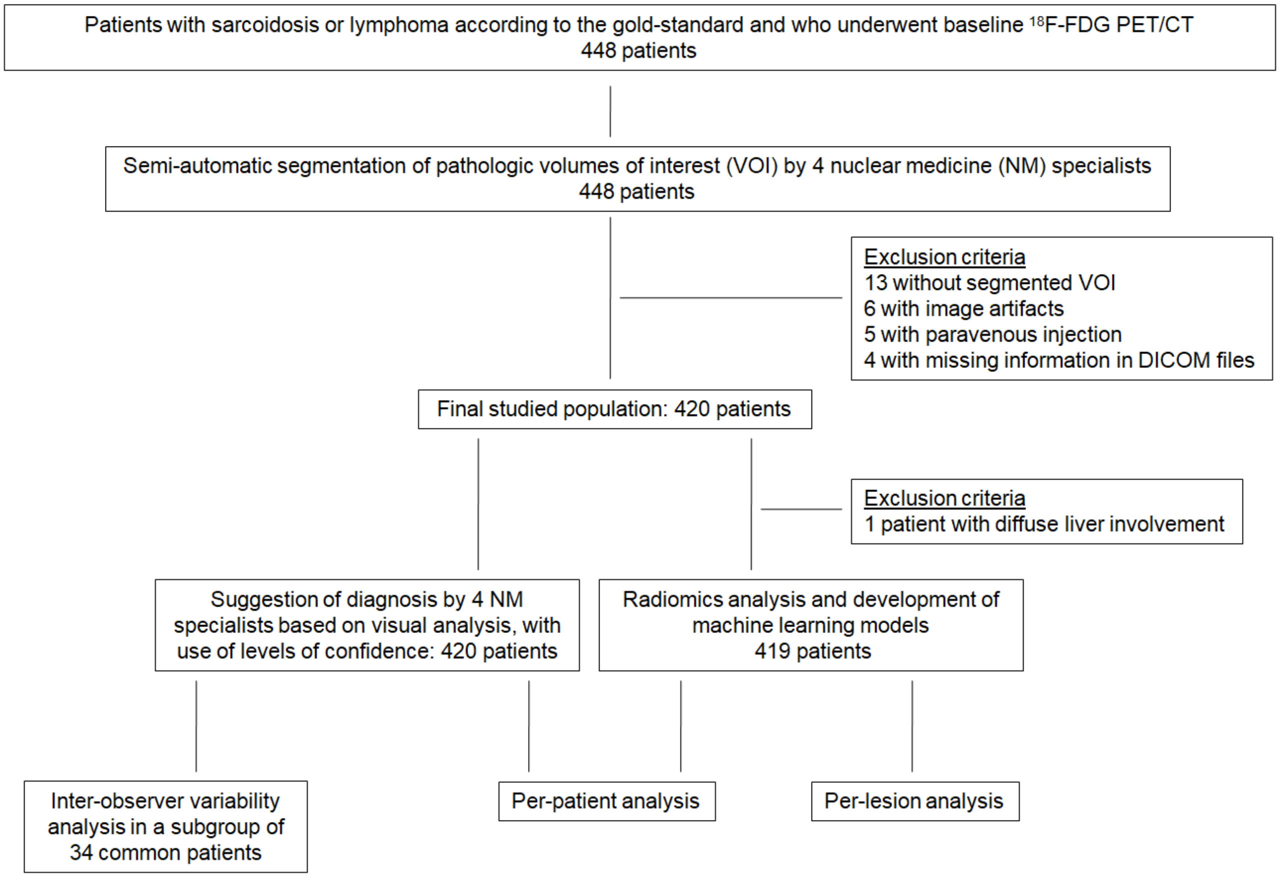

Patients with sarcoidosis or lymphoma according to the gold-standard and who underwent baseline ^{18}F-FDG PET/CT
448 patients
Semi-automatic segmentation of pathologic volumes of interest (VOI) by 4 nuclear medicine (NM) specialists
448 patients
Exclusion criteria
13 without segmented VOI
6 with image artifacts
5 with paravenous injection
4 with missing information in DICOM files
Final studied population: 420 patients
Exclusion criteria
1 patient with diffuse liver involvement
Suggestion of diagnosis by 4 NM specialists based on visual analysis, with use of levels of confidence: 420 patients
Radiomics analysis and development of machine learning models
419 patients
Inter-observer variability analysis in a subgroup of 34 common patients
Per-patient analysis
Per-lesion analysis

Figure 1: Study flowchart

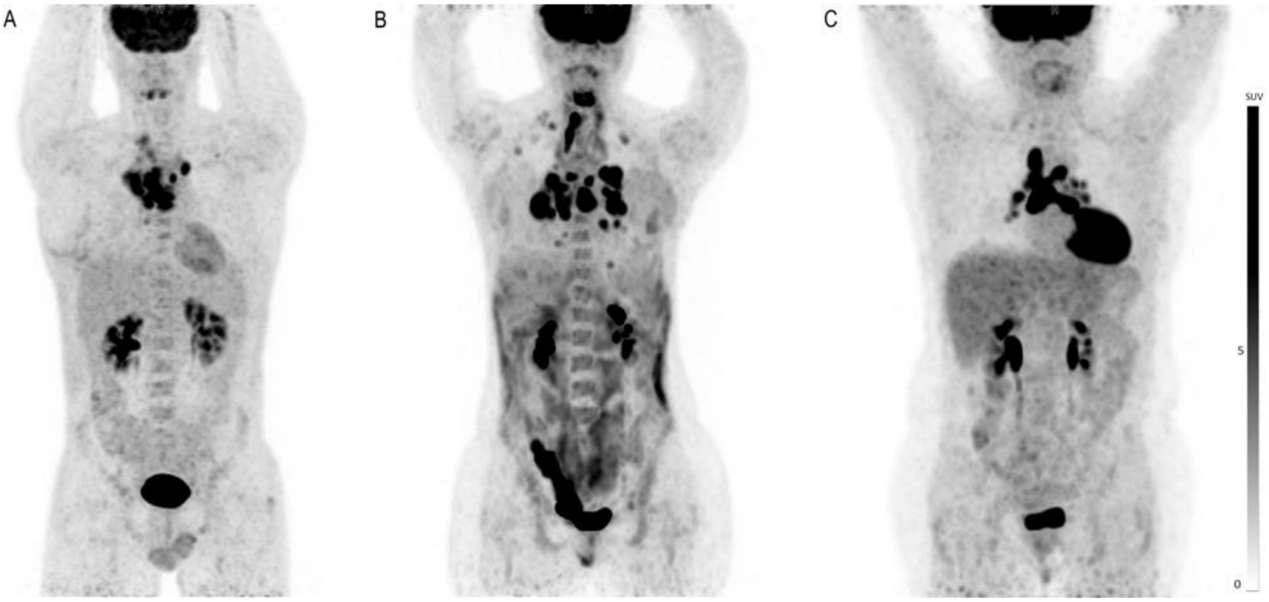


Figure 2: Representative examples of ^{18}F-FDG PET/CT studies of diseases localized to the thorax. A. Diffuse Large B-Cell lymphoma ; B. Hodgkin lymphoma ; C. Sarcoidosis

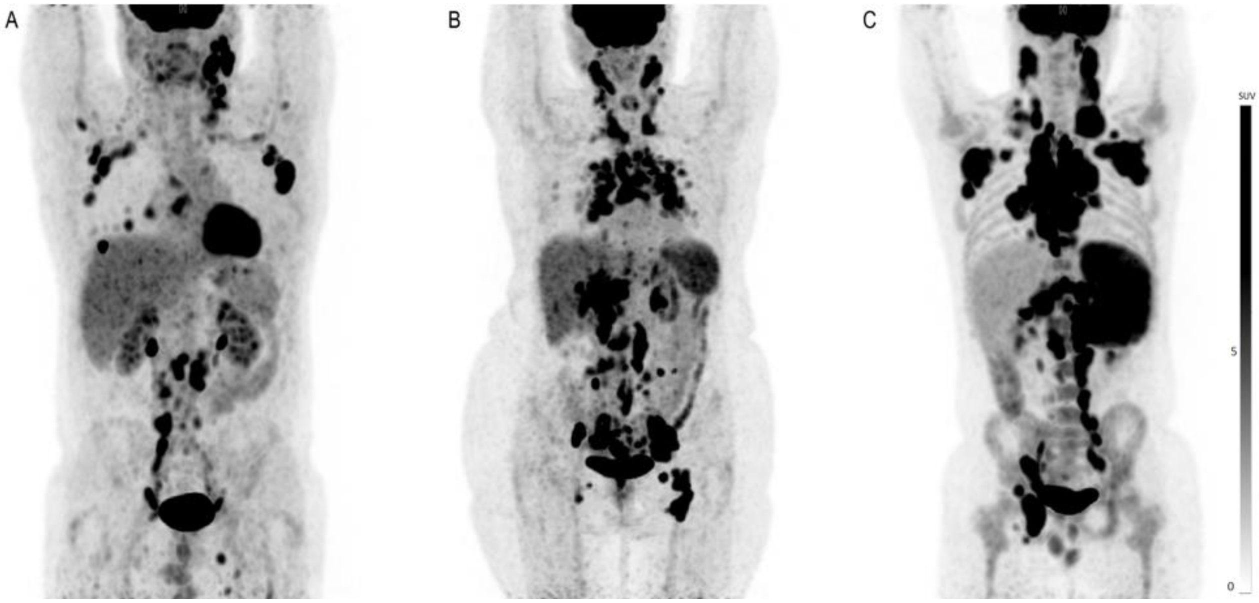


Figure 3: Representative examples of ^{18}F-FDG PET/CT studies of diffuse diseases. A. Diffuse Large B-Cell lymphoma; B. Sarcoidosis; C. Hodgkin lymphoma

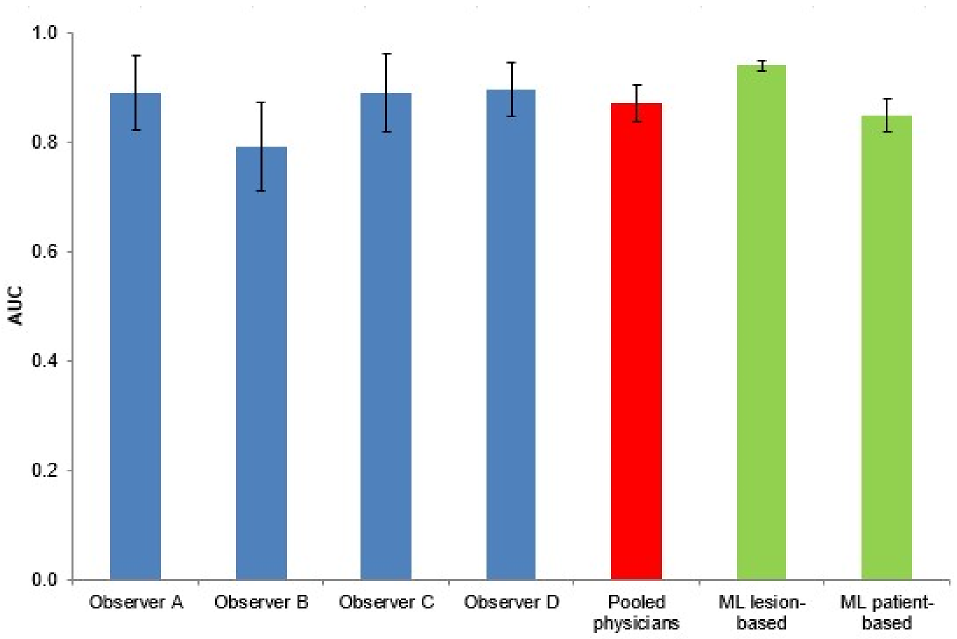


Figure 4: Columns chart illustrating physicians' and machine learning (ML)-radiomics models' performances for the diagnosis of sarcoidosis versus lymphoma. AUC: area under the curve. Vertical lines at the top of each bar represent confidence intervals.

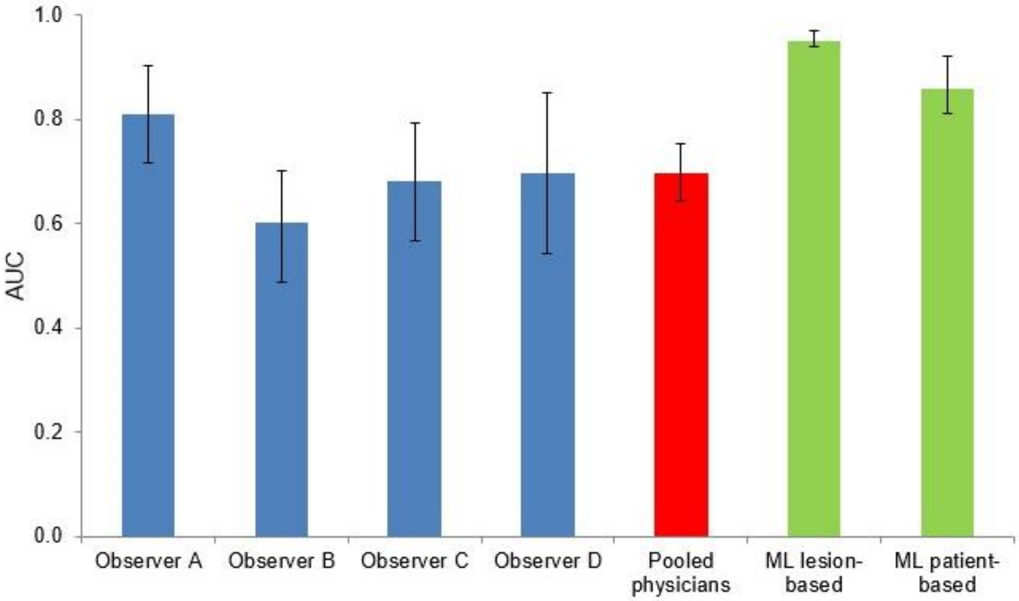
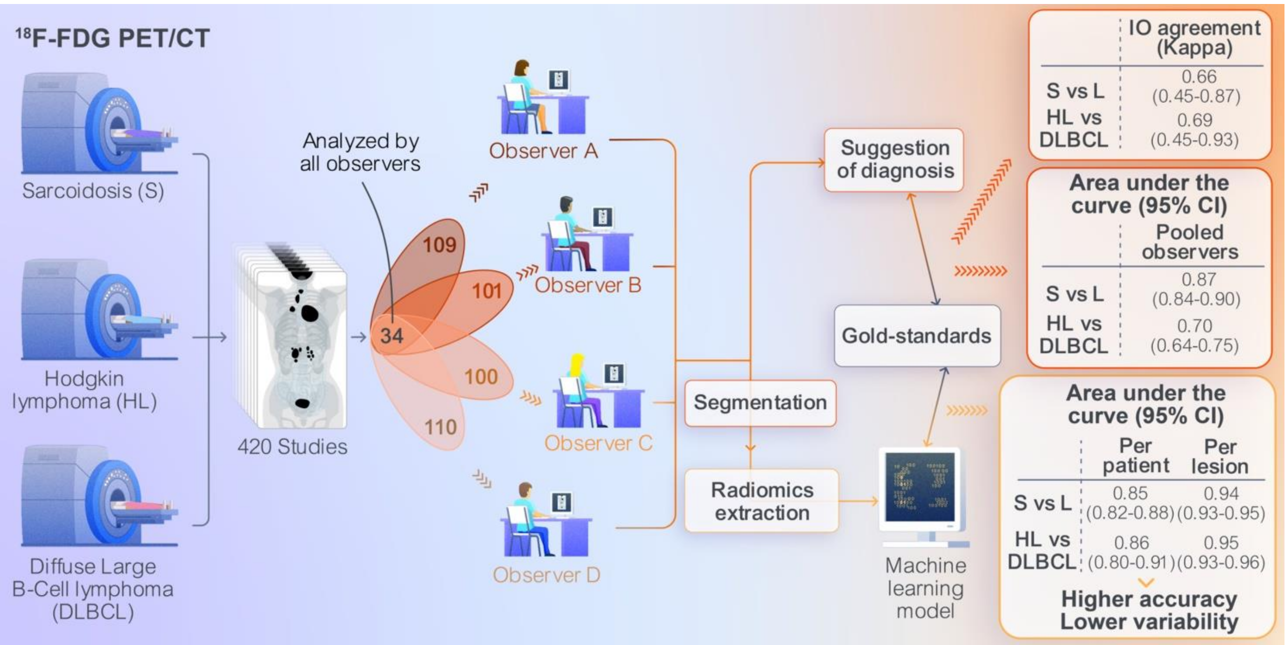


Figure 5: Columns chart illustrating physicians' and machine learning (ML)-radiomics models' performances for the diagnosis of Hodgkin lymphoma versus diffuse large B-cell lymphoma. AUC: area under the curve. Vertical lines at the top of each bar represent confidence intervals.



	IO agreement (Kappa)
S vs L	0.66 (0.45-0.87)
HL vs DLBCL	0.69 (0.45-0.93)

Area under the curve (95% CI)

	Pooled observers
S vs L	0.87 (0.84-0.90)
HL vs DLBCL	0.70 (0.64-0.75)

Area under the curve (95% CI)

	Per patient	Per lesion
S vs L	0.85 (0.82-0.88)	0.94 (0.93-0.95)
HL vs DLBCL	0.86 (0.80-0.91)	0.95 (0.93-0.96)

Higher accuracy
Lower variability

Graphical abstract

Supplemental Table 1. Cohen's kappa for agreement with the gold standard in the diagnosis of BBS vs. Cancer and pairwise comparison of diagnostic performance

Cohen's kappa		**Pairwise comparison**		**Hotelling's T-squared**
Observer A	0.79 (0.55, 1.02)	A and	B	T^2 = 3.92, p=0.056
Observer B	0.45 (0.13, 0.78)		C	T^2 = 0.01, p=0.943
Observer C	0.78 (0.53 - 1.02)		D	T^2 = 2.13, p=0.154
Observer D	0.93 (0.79, 1.07)	B and	C	T^2 = 2.93, p=0.096
			D	T^2 = 8.70, p=0.006
		C and	D	T^2 = 2.21, p=0.146

Supplemental Table 2. Cohen's kappa for agreement with the gold standard in the diagnosis of Hodgkin lymphoma vs. DLBCL and pairwise comparison of diagnostic performance

Cohen's kappa		**Pairwise comparison**		**Hotelling's T-squared**
Observer A	0.70 (0.38, 1.01)	A and	B	T^2 =9.60, p=0.006
Observer B	0.07 (-0.35, 0.49)		C	T^2 = 5.42, p=0.031
Observer C	0.27 (-0.15, 0.69)		D	T^2 = 5.42, p=0.031
Observer D	0.27 (-0.15, 0.69)	B and	C	T^2 = 2.18, p=0.155
			D	T^2 =2.18, p=0.155
		C and	D	(*)

(*) The calculation was not performed due to no variability observed between observers C and D.

Supplemental Table 3

Best radiomic models and its diagnostic performance with95% confidence intervals of ‘Lesion based’ models in the classification of cancer vssarcoidoisis. The best radiomic signature was chosen according to the best mean 5-fold cross validation AUCpr.

Radiomics type	Classifier	Model	Features	AUC	AUCpr	Sensitivity	Specificity	PPV	NPV
OR	RF	FS: RF_Gini; Discretization:FBW (0.5 SUV);	Stats_min;Shape_areaDensityBE;Shape_flatness;Shape_elongation;Shape_volumeDensityBB;Shape_centroidDistance;Shape_areaDensityBB;Shape_volumeDensityBE;Stats_p10;NGLDM_HGLDE;GLCM_maxCorr;IVH_AIRV_80;GLSZM_SAE;NGLDM_DV;IVH_AIRV_90;GLSZM_LISAE;NGTDM_busyness;IVH_AIRV_50;Shape_leastaxislength;Shape_maxDiameter2D2	0.68 (0.66-0.70)	0.78 (0.77-0.80)	0.83 (0.79-0.88)	0.34 (0.27-0.41)	0.68 (0.67-0.70)	0.54 (0.49-0.60)
TLR	RF	FS: RF_Accuracy; Discretization:FBW (0.05 SUV);	Stats_min;IVH_AIRV_90; Stats_p10;GLCM_infoCorr2	0.94 (0.93-0.95)	0.96 (0.95-0.97)	0.92 (0.89-0.94)	0.80 (0.75-0.84)	0.88 (0.86-0.91)	0.85 (0.81-0.89)

Supplemental Table 4

Best radiomic models and its diagnostic performance with95% confidence intervals of ‘Lesion-based’ models in the classification of HLvsDLBCL. The best radiomic signature was chosen according to the best mean 5-fold cross validation AUCpr.

Radiomics type	Classifier	Model	Features	AUC	AUCpr	Sensitivity	Specificity	PPV	NPV
OR	RF	FS: MRMR_Forward_spearman; Discretization:FBW (0.2 SUV);	GLCM_inverseVar;Stats_skewness;Stats_min;IVH_RVRI_20;IH_maxGradI;IVH_RVRI_10;GLCM_maxCorr;IH_entropy;IH_mode;Shape_elongation	0.67 (0.64-0.70)	0.62 (0.58-0.66)	0.57 (0.50-0.64)	0.66 (0.60-0.73)	0.60 (0.56-0.64)	0.64 (0.60-0.67)
TLR	RF	FS:RF_Gini; Discretization:FBW (0.05 SUV);	GLCM_infoCorr2;Stats_min	0.95 (0.93-0.96)	0.95 (0.92-0.96)	0.89 (0.85-0.92)	0.88 (0.84-0.92)	0.87 (0.83-0.90)	0.90 (0.87-0.92)

Supplemental Table 5

Best radiomic models and its diagnostic performance with95% confidence intervals of ‘Patient-based’ models in the classification of cancervssarcoidosis. The best radiomic signature was chosen according to the best mean 5-fold cross validation AUCpr.

Radiomics type	Radiomics merge metric	Classifier	Model	Features	AUC	AUCpr	Sensitivity	Specificity	PPV	NPV
OR	Mean	LR	FS: MRMR_Backward_pearson; Discretization:FBW (0.1 SUV);	Stats_skewness;Stats_min;Stats_kurtosis;Shape_volumeDensityBE;Shape_volumeDensityBB;Shape_sphericity;Shape_spherDisprop;Shape_flatness;Shape_elongation;Shape_compactness3;Shape_compactness2;Shape_asphericity;Shape_areaDensityBE;Shape_areaDensityBB	0.76 (0.72-0.80)	0.79 (0.75-0.83)	0.79 (0.72-0.86)	0.60 (0.50-0.68)	0.75 (0.70-0.78)	0.67 (0.59-0.75)
TLR	Minimum	RF	FS: RF_Accuracy; Discretization:FBN (32 bins)	NGLDM_SM2;Shape_surfVolRatio;Shape_compactness2;GLRLM_RP;IVH_AIRV_70; Shape_volumeDensityBE	0.82 (0.79-0.85)	0.87 (0.82-0.90)	0.79 (0.72-0.86)	0.65 (0.56-0.74)	0.77 (0.73-0.82)	0.68 (0.62-0.76)
OR+ Clinical	Maximum	RF	FS: RF_Accuracy; Discretization: FBW (0.5)	Age;GLCM_correl1;GLSZM_HILAE;NGLDM_HGLDE; Shape_asphericity;Shape_spherDisprop	0.85 (0.81-0.89)	0.87 (0.82-0.91)	0.80 (0.72-0.90)	0.74 (0.65-0.82)	0.82 (0.77-0.87)	0.72 (0.64-0.81)
TLR+ Clinical	Minimum	RF	FS:RF_Accuracy; Discretization: FBN (32 bins)	NGLDM_SM2;Shape_surfVolRatio;Age;Shape_volumeDensityBE;NGLDM_DNN;Shape_sphericity;IVH_AIRV_70;GLDZM_DZNN	0.85 (0.82-0.88)	0.88 (0.84-0.92)	0.84 (0.78-0.90)	0.67 (0.56-0.76)	0.79 (0.74-0.84)	0.74 (0.67-0.83)

Supplemental Table 6

Best radiomic models and its diagnostic performance with95% confidence intervals of ‘Patient-based’ models in the classification of HLvsDLBCL. The best radiomic signature was chosen according to the best mean 5-fold cross validation AUCpr.

Radiomics type	Radiomics merge metric	Classifier	Model	Features	AUC	AUCpr	Sensitivity	Specificity	PPV	NPV
OR	Maximum	NB	FS:MRMR_Forward_pearson; Discretization: FBW (0.5 SUV)	IH_entropy;NGLDM_GLN;IVH_AIRV_70;Shape_compactness2;GLCM_infoCorr1;GLSZM_SZN;GLSZM_SZNN;GLDZM_LDE;IH_minGradI;Stats_max;GLCM_entrop2;IH_maxGrad;NGLDM_SDE;IVH_AVRI_50;IH_medianD;NGLDM_LGLDE;GLCM_maxCorr;IVH_AIRV_80;GLDZM_IN;NGLDM_DNN	0.73 (0.67-0.77)	0.70 (0.64-0.73)	0.71 (0.61 - 0.79)	0.72 (0.64-0.77)	0.76 (0.71-0.81)	0.66 (0.59-0.74)
TLR	Minimum	RF	FS:RF_Accuracy; Discretization: FBW (0.1 SUV)	IVH_RVRI_20;NGLDM_SM2;Shape_surfVolRatio;IVH_AVRI_40;GLSZM_INN;IH_uniformity;GLCM_homogeneity2;GLDZM_INN;GLRLM_GLNN;NGLDM_GLNN;IVH_RVRI_30;GLCM_homogeneity1;GLDZM_SDE;GLSZM_HILAE;NGLDM_LGSDE;Shape_maxDiameter3D;GLSZM_LIE;Shape_volumeDensityBE;NGTDM_coarseness;IVH_AVRI_50	0.67 (0.61-0.72)	0.67 (0.60-0.75)	0.71 (0.61-0.79)	0.56 (0.45-0.68)	0.67 (0.62-0.73)	0.60 (0.52-0.68)
OR+ Clinical	Maximum	NB	FS:MRMR_Forward_pearson; Discretization: FBW (0.5 SUV)	IH_entropy;Age;NGLDM_SM;IVH_AIRV_70;GLCM_infoCorr1;NGLDM_DNN	0.86 (0.80-0.91)	0.87 (0.78-0.91)	0.79 (0.71-0.86)	0.85 (0.73-0.86)	0.87 (0.79-0.89)	0.76 (0.70 - 0.83)

TLR+ Clinical	Maximum	RF	FS: MRMR_Forward_spearman; Discretization: FBW (0.5 SUV)	GLRLM_LRHGE;Age;IH_cov;IVH_TLGRI_60;GLSZM_LAE;Shape_compactness2;GLSZM_SZN;NGLDM_HGLDE;GLCM_inverseVar;GLDZM_IN; GLSZM_SZNN;Sexe	0.79 (0.72-0.85)	0.81 (0.72-0.88)	0.68 (0.57-0.82)	0.74 (0.64-0.82)	0.77 (0.71-0.84)	0.65 (0.57-0.76)