18F-FDG PET maximum intensity projections and artificial intelligence: a win-win combination to easily measure prognostic biomarkers in DLBCL patients

Kibrom Berihu Girum, Louis Rebaud, Anne-Ségolène Cottereau, Michel Meignan, Jérôme Clerc, Laetitia Vercellino, Olivier Casasnovas, Franck Morschhauser, Catherine Thieblemont, Irène Buvat

Journal of Nuclear Medicine, published on June 16, 2022 as doi:10.2967/jnumed.121.263501

^{18}F-FDG PET maximum intensity projections and artificial intelligence: a win-win combination to easily measure prognostic biomarkers in DLBCL patients

Kibrom B. Girum[1], Louis Rebaud[1,2], Anne-Ségolène Cottereau[1,3], Michel Meignan[4], Jérôme Clerc[3], Laetitia Vercellino[5], Olivier Casasnovas[6], Franck Morschhauser[7], Catherine Thieblemont[8], Irène Buvat[1]

[1]LITO laboratory, U1288 Inserm, Institut Curie, University Paris-Saclay, Orsay, France

[2]Research and Clinical Collaborations, Siemens Medical Solutions USA, 810 Innovation Dr, Knoxville, TN 37932, United states

[3]Department of Nuclear Medicine, Cochin Hospital, AP-HP, Paris Descartes University, Paris, France

[4]Lysa Imaging, Henri Mondor University Hospital, AP-HP, University Paris East, Créteil, France

[5]Department of Nuclear Medicine, Saint-Louis Hospital, AP-HP, Paris, France

[6]Department of Hematology, University Hospital of Dijon, Dijon, France

[7]Department of Hematology, Claude Huriez hospital, University Lille, EA 7365, Research Group on Injectable Forms and Associated Technologies, Lille, France

[8]Department of Hematology, Saint Louis Hospital, AP-HP, Paris, France

Corresponding author:

Dr. Irène Buvat

LITO laboratory, U1288 Inserm, Institut Curie, University Paris Saclay, Orsay, France

Rue Henri Becquerel, CS 90030, 91401 ORSAY Cedex, France

Mail: irene.buvat@u-psud.fr

Tel: +3362392164

ORCID iD: 0000-0002-7053-6471

First author:

Kibrom B. Girum (Ph.D.)

Postdoctoral researcher

LITO laboratory, UMR 1288 Inserm, Institut Curie, University Paris Saclay, Orsay, France

Mail: kibrom.girum@curie.fr

ORCID iD: 0000-0003-2511-0225

Funding: The REMARC and LNH073B clinical studies and analyses were sponsored by the Lymphoma Academic Research Organization (LYSARC) of France. This study has received funding from ANR (ANR-19-SYME-0005-03).

Word count: 4870

Running title: PET MIP prognostic biomarkers in DLBCL

ABSTRACT

Background: Total metabolic tumor volume (TMTV) and tumor dissemination (Dmax) calculated from baseline ^{18}F-FDG PET/CT images are prognostic biomarkers in Diffuse Large B-cell lymphoma (DLBCL) patients. Yet, their automated calculation remains challenging.

Purpose: To investigate whether TMTV and Dmax features could be replaced by surrogate features automatically calculated using an artificial intelligence (AI) algorithm from only two maximum intensity projections (MIP) of the whole-body ^{18}F-FDG PET images.

Methods: Two cohorts of DLBCL patients from the REMARC (NCT01122472) and LNH073B (NCT00498043) trials were retrospectively analyzed. Experts delineated lymphoma lesions from the baseline whole-body ^{18}F-FDG PET/CT images, from which TMTV and Dmax were measured. Coronal and sagittal MIP images and associated 2D reference lesion masks were calculated. An AI algorithm was trained on the REMARC MIP data to segment lymphoma regions. The AI algorithm was then used to estimate surrogate TMTV (sTMTV) and surrogate Dmax (sDmax) on both datasets. The ability of the original and surrogate TMTV and Dmax to stratify patients was compared.

Results: 382 patients (mean age, 62.1 years ±13.4 [standard deviation]; 207 men) were evaluated. sTMTV was highly correlated with TMTV for REMARC and LNH073B datasets (Spearman r=0.878 and r=0.752 respectively), and so were sDmax and Dmax (r=0.709 and r=0.714 respectively). The hazard ratios (HR) for progression free survival of volume and MIP-based features derived using AI were similar, e.g., TMTV: 11.24 (95% confidence interval (CI): 2.10-46.20), sTMTV: 11.81 (95% CI: 3.29-31.77), and Dmax: 9.0 (95% CI: 2.53-23.63), sDmax: 12.49 (95% CI: 3.42-34.50).

Conclusion: Surrogate TMTV and Dmax calculated from only 2 PET MIP images are prognostic biomarkers in DLBCL patients and can be automatically estimated using an AI algorithm.

Keywords: Artificial intelligence; DLBCL; ^{18}F FDG PET/CT; dissemination; metabolic tumor volume

INTRODUCTION

Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma. In clinical practice, ^{18}F-FDG PET/CT is a standard-of-care for staging and assessing response in DLBCL patients (*1*). The prognostic value of the total metabolically active tumor volume (TMTV) measured from the whole-body ^{18}F-FDG PET/CT images performed before treatment has been widely demonstrated in lymphoma, especially in DLBCL (*2*–*6*). The disease dissemination reflected by the largest distance between two lesions in the baseline whole-body ^{18}F-FDG PET/CT image (Dmax) has been recently shown to be a complementary early prognostic factor compared to TMTV (*7*,*8*). TMTV and Dmax calculations require tumor volume delineation over the whole-body three-dimensional (3D) ^{18}F-FDG PET/CT images, which is prone to observer-variability and complicates the use of these quantitative features in clinical routine.

To address this problem, automated lesion segmentation approaches using convolutional neural networks (CNN) have been proposed (*9*,*10*). These methods require high computational resources to be developed but have shown promising results, despite missing small lesions (*7*). Yet, results from CNN still need to be validated and adjusted by an expert before using them for further analysis (*7*,*8*). This implies a thorough visual analysis of all 3D ^{18}F-FDG PET/CT images and delineation of the lesions missed by the algorithm. Consequently, developing a pipeline that would speed-up this checking/adjustment process is highly desirable in clinical practice.

Nuclear medicine physicians commonly use two-dimensional (2D) PET Maximum Intensity Projection (MIP) views for visual interpretation as a synthetic representation of the 3D distribution of the tracer over the whole-body. However, to the best of our knowledge, the prognostic value of PET parameters extracted from 2D MIP has never been explored. We hypothesized that tumor burden and spread could be automatically evaluated from only two PET MIP images corresponding to coronal and sagittal views. This would have two advantages: first, result checking and adjustment would be faster from two MIP views compared to the whole-body 3D ^{18}F-FDG PET/CT images, typically including more than 200 transaxial slices. Second, a deep learning model for segmenting MIP images would involve much fewer parameters

than when segmenting the whole-body 3D ^{18}F-FDG PET images. It is less computationally expensive and might require less data for training.

The purpose of this study was to investigate whether TMTV and Dmax biomarkers could be replaced by surrogate biomarkers automatically calculated using an artificial intelligence (AI) algorithm from only two MIP of the whole-body ^{18}F-FDG PET images. We then determined the prognostic values of the surrogate biomarkers in terms of progression-free survival (PFS) and overall survival (OS).

MATERIALS AND METHODS

Patient Cohorts

The study population included DLBCL patients who had a baseline (before treatment initiation) PET/CT scan from two independent trials: REMARC (NCT01122472) and LNH073B (NCT00498043). The characteristics of these cohorts have been published elsewhere ((*6*) for REMARC, (*11*) for LNH073B) (*12*). PFS and OS as defined following the revised National Cancer Institute criteria (*13*) were recorded. Flow diagrams for the datasets and the study design are summarized in Figure 1. All data were anonymized before analysis. The institutional review board approval, including ancillary studies, was obtained for the two trials, and all patients provided written informed consent. The demographics and staging of the patients used for the survival analysis are summarized in Table 1.

Measurements of Reference TMTV and Dmax

For the REMARC cohort, the lymphoma regions were identified in the 3D PET images as described in (*6*,*14*), while the LNH073B lesions were segmented as explained in (*7*). In all cohorts, physicians removed the regions corresponding to physiological uptakes and added pathological regions missed by the algorithm. The supplementary material (section A) provides the details. Expert-validated 3D lymphoma regions were used to compute the reference TMTV and Dmax (based on the centroid of the lymphoma regions), as shown in Figure 1(B) (*8*).

Calculation of the PET MIP Images and 2D Reference Lymphoma Regions

For each patient whole-body 3D ^{18}F-FDG PET images and associated 3D lymphoma regions, two 2D MIP views and associated 2D lymphoma regions were calculated (Figure 2). The 3D PET image was projected in the coronal and sagittal directions, 90° apart (Figure 2), setting each pixel value of the projection to the maximum intensity observed along the ray normal to the plane of projection. Similarly, MIP of the expert-validated 3D lymphoma regions were calculated, resulting in binary images of 2D lymphoma regions (Figure 2), hereafter called MIP_masks. As described in the following section, these MIP_masks were then used as a reference output to train a CNN-based fully automatic lymphoma segmentation method.

Fully Automatic Lymphoma Segmentation on PET MIP Images

Deep Learning Model Inputs and Architectures. To automatically segment the lymphoma lesions from the sagittal and coronal PET MIP images, we adapted a previously published supervised 2D deep learning model (*15*). The sagittal and coronal PET MIPs were independent input images during training. The corresponding MIP_mask was the output image. The deep learning model was trained to transform a given sagittal or coronal PET MIP image to the corresponding MIP_mask with pixels of lymphoma regions set to one and pixels of the non-lymphoma regions set to zero.

Training, Validation, and Testing Configurations. First, using the REMARC cohort (298 patients), a five-fold cross-validation technique was used to train and evaluate the model. Patients were randomly split into five groups, and then five models were trained on 80% of the population and the remaining 20% was used for validation. The detailed network architecture (*15*,*16*) and the training procedures are fully described in the supplementary material (section B, Supplemental Figure 1) (*17*), following the CLAIM guidelines (*18*) and SNMMI AI Task force recommendations (*19*). The deep learning model will be publicly available upon publication [GitHub, Anonymous].

Secondly, we tested the model trained from the REMARC cohort (298 patients) on the independent LNH073B cohort (174 patients) to characterize its generalizability and robustness. The REMARC and LNH073B cohorts were acquired from two different trials. The REMARC study data was a double-blind,

international, multicenter, randomized phase III study, which started inclusion in 2010. In contrast, the LNH073B data was a prospective multicenter, randomized phase II study that started including patients in 2007.

Calculation of Surrogate TMTV and Surrogate Dmax

The surrogate TMTV (sTMTV) and Dmax (sDmax) were defined and computed from the MIP_masks automatically segmented from the coronal and sagittal PET MIP images using the deep learning method.

Tumor Burden Analysis. To characterize tumor burden, we defined a surrogate tumor volume sTMTV from the MIP_mask as the number of pixels belonging to the tumor regions in MIP_mask multiplied by the pixel area. For a given patient, sTMTV was calculated from the coronal and the sagittal MIP_masks as $sTMTV = sTMTV_{coronal} + sTMTV_{sagittal}$.

Tumor Dissemination Analysis. The spread of the disease was analyzed by estimating the largest distance between the tumor pixels belonging to the MIP_mask, using a new robust largest distance estimation approach. First, we separately calculated the sum of pixels along the columns and the rows of MIP_mask, yielding x and y profiles (Supplemental Figure 2). Second, in each of these two profiles, the distances between the 2% percentile and the 98% percentiles ($x_{2\%}$ and $x_{98\%}$ in the x profiles, $y_{2\%}$ and $y_{98\%}$ in the y profiles) were calculated, yielding $(x_{98\%} - x_{2\%})$ and $(y_{98\%} - y_{2\%})$, respectively. These percentiles were chosen to improve the robustness of the calculation to outliers. The largest distance was defined as $sDmax_{sagittal/coronal} = (x_{98\%} - x_{2\%}) + (y_{98\%} - y_{2\%})$. For a given patient, the surrogate tumor dissemination sDmax was the sum of the coronal and sagittal disseminations using $sDmax = sDmax_{sagittal} + sDmax_{coronal}$.

Statistical Analysis

Using the MIP_masks obtained from the expert-delineated 3D lymphoma regions (Figure 2) as a reference, CNN's segmentation performance was evaluated using the Dice score, sensitivity, and specificity. The difference between the CNN-based segmentation results and the expert-delineated 3D

lymphoma regions were quantified using Wilcoxon statistical tests. Univariate and multivariate survival analyses were performed. For all biomarkers, we calculated a time-dependent area under the receiver operating characteristics curve (AUC) (*20*). Bootstrap resampling analysis was performed to associate confidence intervals to the Cox model hazard ratio and the time-dependent AUC. See the supplementary material (section C) for details. Test results were considered statistically significant if the two-sided P-value was <0.05.

RESULTS

A total of 475 patients from two different cohorts were included in this study, of which 93 patients were excluded from the biomarker and survival analysis because the provided baseline ^{18}F-FDG PET/CT images were not suitable to analyze all biomarkers (no PET segmentation by an expert or less than 2 lesions). Summary statistics of patients are presented in Table 1.

Lymphoma Segmentation

The performance of the proposed segmentation method was evaluated patient-wise. The CNN segmentation method achieved a 0.80 median Dice score (interquartile range [IQR]: 0.63-0.89), 80.7% (IQR: 64.5%-91.3%) sensitivity, and 99.7% (IQR: 99.4%-0.99.9%) specificity on the REMARC cohort. On the testing 174 LNH073B patients, the CNN yielded a 0.86 median Dice score (IQR: 0.77-0.92), 87.9% (IQR: 74.9.0%-94.4%) sensitivity, and 99.7% (IQR: 99.4%-99.8%) specificity. In the LNH073B data, the CNN yielded a mean Dice score of 0.80 ± 0.17 (mean ± SD) on the coronal view and 0.79 ± 0.17 on the sagittal view. Figure 3 shows segmentation result examples from experts (MIP_masks) and CNN. See Supplemental Figure 3 for more segmentation results. The Dice score was not significantly different (P >0.05) between the coronal and sagittal views, both for the REMARC and LNH073B cohorts (p>0.05).

In both cohorts, there was a significant correlation between ranked TMTV and Dmax values and the associated surrogate values obtained using CNN. For REMARC, TMTV was correlated with sTMTV (Spearman r = 0.878, p<0.001), and Dmax was correlated with sDmax (r = 0.709, p<0.001). Out of 144 patients who had TMTV greater than the median TMTV (242 cm^3), 121 (84.02%) patients had also sTMTV

greater than the median sTMTV (174 .24 cm^2). 144 patients had Dmax greater than the median Dmax (44.8 cm), and 113 (78.5%) of these patients also had sDmax greater than the median sDmax (98.0 cm).

For LNH073B, TMTV was correlated with sTMTV (r =0.752, p<0.001), and Dmax was correlated with sDmax (r = 0.714, p<0.001). Out of 48 patients who had TMTV greater than the median TMTV (375 cm^3), 42 (87.5%) patients had also sTMTV greater than the median sTMTV (307.2 cm^2). 48 patients had Dmax greater than the median Dmax (44.1 cm), and 39 (81.3%) of these patients also had sDmax greater than the median sDmax (116.4 cm). Table 2 shows the descriptive statistics for the surrogate PET features.

Survival Analysis

The time-dependent AUC and hazard ratios (HR) with 95% confidence interval of the metabolic tumor volume and tumor spread are shown in Table 3 for the REMARC and LNH073B data. All PET features extracted from the baseline 3D ^{18}F-FDG PET/CT images and using AI (sTMTV and sDmax) were significant prognosticators of the PFS and OS.

Combining TMTV and Dmax (or their surrogates), three risk categories could be differentiated in the REMARC data (Figure 4): using the 3D features, category 1 corresponded to low TMTV (≤ 222 cm^3) and low Dmax (≤ 59 cm) (low risk, n=108); category 2 corresponded to either high Dmax or high TMTV (intermediate risk, n=112); category 3 corresponded to both high Dmax and high TMTV (high risk, n=67). This stratification was similar when using the MIP-features-based categories using AI (Figure 4). The accuracy of the CNN-based classification into three categories with respect to the 3D-biomarkers-based classification was 71.4%.

In the LNH073B cohort, combining TMTV and Dmax (or their surrogates), three risk categories could be differentiated (Figure 5). Using the 3D features, category 1 was defined as low TMTV (≤ 468 cm^3) and low Dmax (≤ 60 cm) (n=45); category 2 corresponded to either high Dmax or high TMTV (n=37); category 3 corresponded to both high Dmax and high TMTV (n=13). Out of the 13 patients classified as high risk, 9 (69.2%) patients had less than 4-years of OS, and 10 (76.9%) patients had less than 4-years

of PFS. This stratification was similar when using the CNN-based results. The sTMTV cut-off value was 376 cm^2, the sDmax cut-off value was 122 cm. There were 38 patients in category 1, 35 in category 2, and 22 in category 3. Out of the 22 patients classified as a high risk, 19 (77.3%) patients had less than 4-years of OS, and 19 (86.4%) patients had less than 4-years of PFS. The accuracy of the AI-based classification into three categories with respect to the 3D-biomarkers-based classification was 64.2%. All patients classified as high risk using the 3D biomarkers were also classified as high risk using the CNN, except one patient who had an OS of 36.6 months. Out of the nine patients classified as high risk when using the CNN but not when using the 3D biomarkers, 8 (88.9%) patients had less than 4-years of OS, and the remaining one (11.1%) patient had 21.95 and 57.99 months of PFS and OS respectively.

In Supplemental Figure 4, the confusion matrices show the agreement between the 3D-based biomarkers and the surrogate MIP biomarkers in the LNH073B data. The percentage of the data classified into high, low, and intermediate risk is also shown. Using a classification in two groups based on one biomarker only (either tumor burden or dissemination biomarkers), the AI-based classification had a 79% accuracy compared to the 3D-based classification.

DISCUSSION

We developed and evaluated a new framework to calculate surrogate metabolic tumor volume (sTMTV) and surrogate tumor dissemination (the largest distance between lymphoma sites) (sDmax) features from 2D PET MIP images. The motivation for considering tumor delineation on 2D MIP views instead of the 3D volume was twofold: first, checking lymphoma regions on 2D PET MIP images is much faster than on the 3D ^{18}F-FDG PET/CT volumes. Second, training an automated AI tumor segmentation algorithm is easier in 2D than in 3D from a practical point of view (fewer parameters to be tuned, less data to be used for training, and less computational cost). We thus investigated the prognostic values of these surrogate biomarkers using two independent retrospective cohorts of DLBCL patients with baseline ^{18}F-FDG PET/CT. Characterizing tumor burden and its dissemination was feasible using features extracted from the 2D PET MIP images. TMTV and Dmax were highly correlated with sTMTV and sDmax, respectively.

Developing automatic and robust lymphoma segmentation methods on PET MIP images could cost less data and less computational resources than when using the whole-body ^{18}F-FDG PET images. It could allow AI experts to quickly investigate appropriate segmentation approaches to tackle the challenging lymphoma segmentation task and reduce inter-center and inter-expert variations in lymphoma delineation. Experts can validate and correct, if necessary, AI results on 2D MIP images easier and faster than on their 3D volume counterparts. We also showed that a convolutional neural network (CNN) could segment lymphoma lesions fully automatically from the given 2D PET MIP image with high accuracy compared with expert readers. This result was confirmed on the independent LNH073B cohort. The proposed CNN-based segmented regions enabled features extraction with predictive values comparable to when these features are calculated from the areas delineated by experts in the 3D image. The main strength of this work was that we validated our findings using an external cohort from a different retrospective trial. However, training the proposed deep learning model from an increased training sample size, preferably from different centers and acquisition parameters, might further improve its performance. No correlations were observed between the segmentation errors made by the model and lesion size. Previous lymphoma segmentation methods used the whole-body ^{18}F-FDG PET/CT images (*9*,*10*). Most of these methods involved complex preprocessing, CT and PET image alignment, and did not investigate whether both TMTV and Dmax remained good prognosticators when calculated from the automated segmentation. Recent studies have also demonstrated that CNN-based results need corrections by experts (*7*,*8*). Correction of results on 3D volume could be time-consuming, observer-dependent, and difficult. In contrast, corrections, and validations (if necessary) could be easier and faster on 2D PET MIP images, allowing easy use of these features in clinical routine.

Interestingly, the surrogate biomarkers automatically calculated using AI (sTMTV and sDmax) had strong prognostic values regarding PFS and OS, comparable to the prognostic importance of TMTV and Dmax obtained from the 3D volumes. The classification of patients into the three risk groups using the 3D TMTV and Dmax agreed with the patient's classification based on the 2D sTMTV and sDmax (71.4% and 64.2% respectively in REMARC and LNH037 cohorts). Patients classified as high-risk using 3D-based biomarkers and low-risk (or vice versa) using 2D-based biomarkers had values close to the cut-off values.

Visual assessment of the segmentation results suggested that the 2D-based biomarkers tend to perform well compared to the 3D-based biomarkers when the patient had lesions spread over the body and performed less well when the patient had a large bulky lesion.

In this work, we defined and calculated the surrogate biomarkers from both the coronal and sagittal PET MIPs. However, experiments showed that characterizing the lymphoma disease using sTMTV and sDmax calculated from either coronal or sagittal also had good predictive values, comparable to these features obtained from 3D volumes. The same cut-off values were used to analyze the PFS and OS. The cohorts were from two independent studies with varying stages of cancer (Table 1). Thus the (s)TMTV cut-off values were different between the two cohorts. Interestingly, the cut-off values to characterize the lesion dissemination (Dmax and sDmax) in DLBCL patients on baseline PET images were almost identical on the independent cohorts. Dmax and sDmax were defined empirically, yet a recent study has shown that the distance between lesions calculated using different distance measurement methods (namely Euclidean, Manhattan, and Chebyshev) in 3D yielded similar results in predictions of the outcome (*21*).

Our study has limitations. Although we validated the CNN on two independent retrospective studies, validating the proposed CNN in larger multicenter cohorts will be required to develop it into a clinical tool. In addition, although the CNN results can be easily visually checked, they should ideally be provided with a confidence level, that could be turned into a confidence associated with the risk classification.

CONCLUSION

In this study, we introduced biomarkers extracted from PET MIP as surrogates of the total metabolic tumor burden and tumor dissemination. To our knowledge, this is the first study showing that PET parameters extracted from 2D MIP images are predictive of outcome in a large series of patients with DLBCL, with results comparable to these features calculated from the 3D ^{18}F-FDG PET/CT images. We demonstrated that surrogate TMTV and Dmax calculated from lymphoma regions automatically delineated on PET MIP images using artificial intelligence have strong prognostic values in stratifying patients with

DLBCL. This result might considerably facilitate the calculation and usage of these features in clinical practices.

DISCLOSURE

Kibrom B. Girum and Irène Buvat disclosed a research grant given to the Institut Curie by ANR (ANR-19-SYME-0005-03). Louis Rebaud disclosed employment by Siemens Medical Solutions. No other potential conflicts of interest relevant to this article exist.

KEY POINTS

QUESTION: Are surrogate tumor burden and dissemination features calculated from PET maximum intensity projection (MIP) images prognostic biomarkers in diffuse large B-cell lymphoma (DLBCL) patients and can they be automatically measured using an AI?

PERTINENT FINDINGS: Surrogate total metabolically active tumor volume (sTMTV) and dissemination feature (sDmax) calculated from MIP of whole-body ^{18}F-FDG PET images are predictive of progression-free survival (PFS) and overall survival (OS) in DLBCL patients from two independent cohorts. A convolutional neural network (CNN) could segment lymphoma lesions from 2D PET MIP images automatically and the resulting CNN-based sTMTV and sDmax estimates were predictive of PFS and OS in two independent cohorts.

IMPLICATIONS FOR PATIENT CARE: Surrogate tumor burden and dissemination features automatically calculated using AI from only two PET MIP images are prognostic biomarkers in DLBCL patients.

REFERENCES

1. 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.
2. Mikhaeel NG, Smith D, Dunn JT, et al. Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. *Eur J Nucl Med Mol Imaging*. 2016;43:1209-1219.
3. Cottereau AS, Lanic H, Mareschal S, et al. Molecular profile and FDG-PET/CT total metabolic tumor volume improve risk classification at diagnosis for patients with diffuse large B-Cell lymphoma. *Clin Cancer Res*. 2016;22:3801-3809.
4. Kostakoglu L, Martelli M, Sehn LH, et al. Baseline PET-derived metabolic tumor volume metrics predict progression-free and overall survival in DLBCL after first-line treatment: Results from the phase 3 GOYA study. *Blood*. 2017;130:824-824.
5. Schmitz C, Hüttmann A, Müller SP, et al. Dynamic risk assessment based on positron emission tomography scanning in diffuse large B-cell lymphoma: Post-hoc analysis from the PETAL trial. *Eur J Cancer*. 2020;124:25-36.
6. Vercellino L, Cottereau AS, Casasnovas O, et al. High total metabolic tumor volume at baseline predicts survival independent of response to therapy. *Blood*. 2020;135:1396-1405.
7. Cottereau A-S, Nioche C, Dirand A-S, et al. 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med*. 2020;61:40-45.
8. Cottereau A-S, Meignan M, Nioche C, et al. Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†. *Ann Oncol*. 2021;32:404-411.
9. Sibille L, Seifert R, Avramovic N, et al. 18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*. 2020;294:445-452.
10. Blanc-Durand P, Jégou S, Kanoun S, et al. Fully automatic segmentation of diffuse large B cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network. *Eur J Nucl Med Mol Imaging*. 2021;48:1362-1370.
11. Casasnovas R-O, Ysebaert L, Thieblemont C, et al. FDG-PET–driven consolidation strategy in

diffuse large B-cell lymphoma: final results of a randomized phase 2 study. *Blood*. 2017;130:1315-1326.

12. Nioche C, Orlhac F, Boughdad S, et al. Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res*. 2018;78:4786-4789.
13. Cheson BD, Pfistner B, Juweid ME, et al. Revised response criteria for malignant lymphoma. *J Clin Oncol*. 2007;25:579-586.
14. Capobianco N, Meignan M, Cottereau A-S, et al. Deep-learning 18 F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-Cell lymphoma. *J Nucl Med*. 2021;62:30-36.
15. Girum KB, Crehange G, Lalande A. Learning with context feedback loop for robust medical image segmentation. *IEEE Trans Med Imaging*. 2021;40:1542-1554.
16. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:770-778.
17. Kingma DP, Ba J. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc*. December 2014:1-15.
18. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol Artif Intell*. 2020;2:e200029.
19. Bradshaw TJ, Boellaard R, Dutta J, et al. Nuclear medicine and artificial intelligence: Best practices for algorithm development. *J Nucl Med*. 2021:jnumed.121.262567. In Press.
20. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56:337-344.
21. Cottereau A-S, Meignan M, Nioche C, et al. New approaches in characterization of lesions dissemination in DLBCL patients on baseline PET/CT. *Cancers (Basel)*. 2021;13:3998.

Figures and Tables

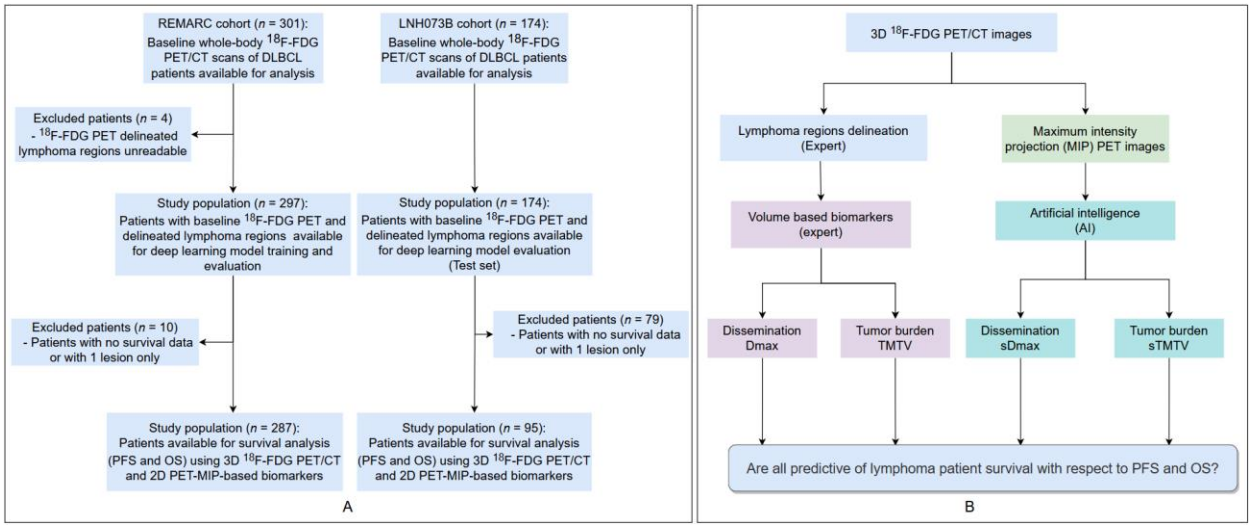


FIGURE 1. A) Study flowchart, B) Study design. FDG=Fluorodeoxyglucose, TMTV = total metabolic tumor volume, Dmax: maximum distance between two lesions, sTMTV: surrogate TMTV automatically calculated using AI, sDmax: surrogate Dmax automatically calculated using AI, PFS: progression-free survival, OS: overall survival.

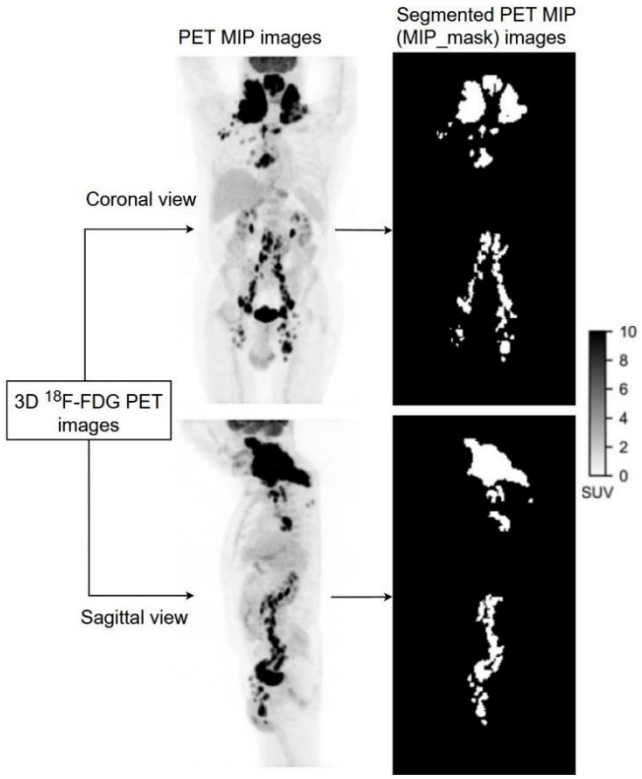


FIGURE 2. An example of ^{18}F-FDG PET MIP images (left) and associated lymphoma regions (right) based on the expert delineation of the 3D ^{18}F-FDG PET images.

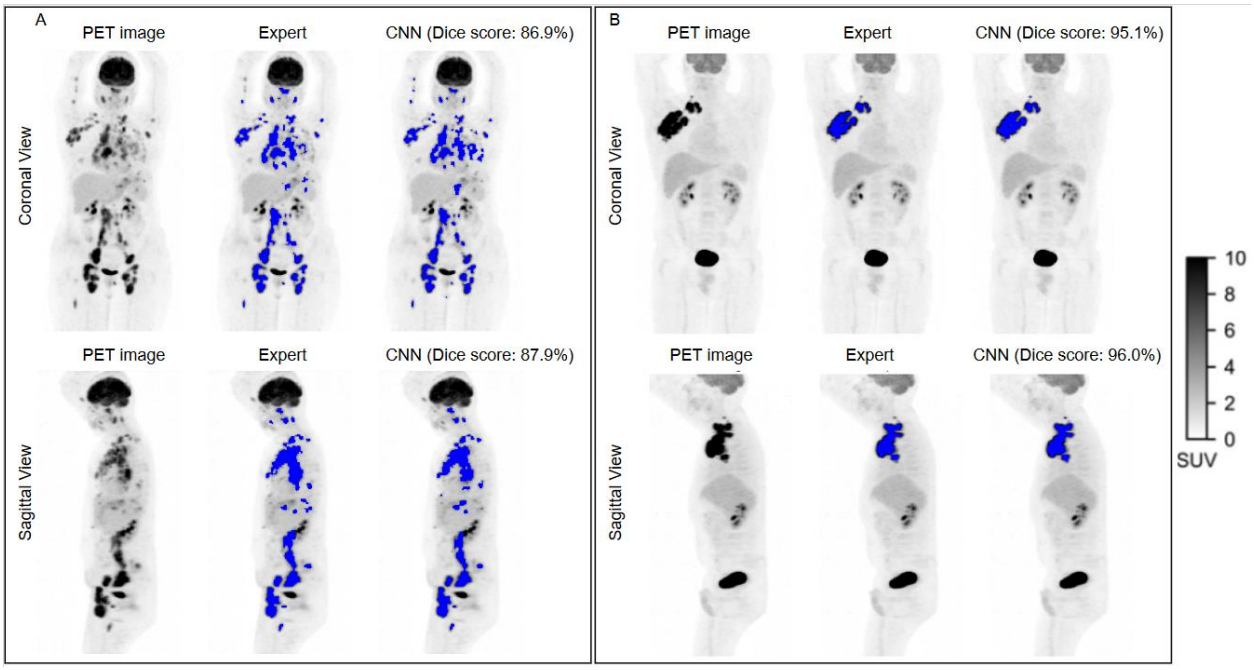


FIGURE 3. ^{18}F-FDG PET MIP images and segmentation results (blue color overlapped over the PET MIP images) by experts (MIP_masks) and by the CNN for four patients: (A) from the REMARC cohort, and (B) from the LNH073B cohort.

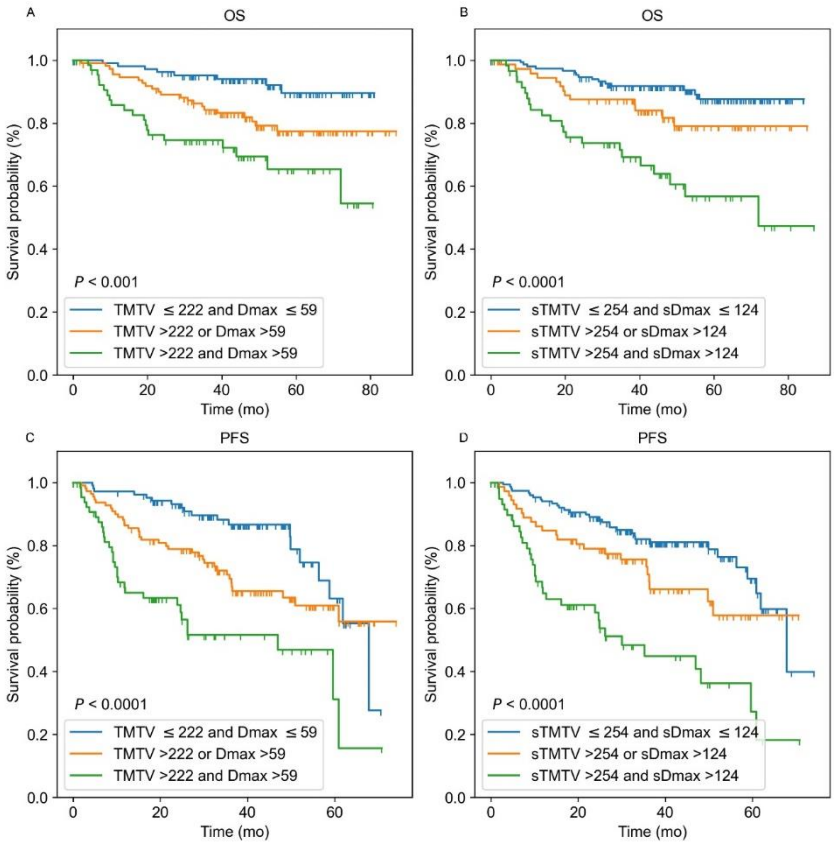


FIGURE 4. Kaplan-Meier estimates of overall survival (OS) and progression-free survival (FPS) on the REMARC cohort according to 3D ^{18}F-FDG PET/CT image-based features TMTV (cm^3) and Dmax (cm) (A, C), and according to PET MIP image-based features (sTMTV (cm^2) and sDmax (cm)) estimated from AI (B, D).

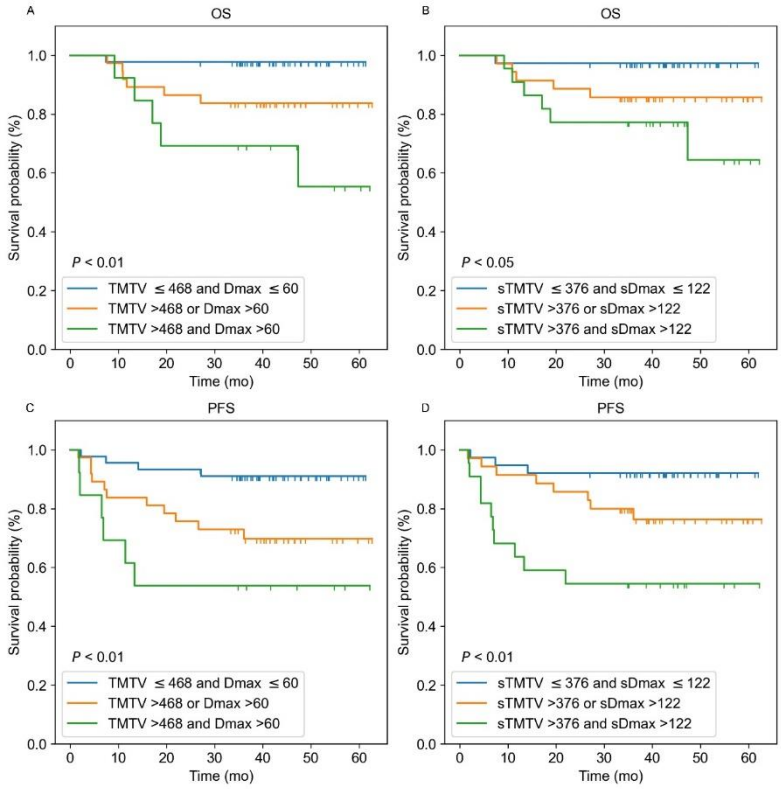


FIGURE 5. Kaplan-Meier estimates of overall survival (OS) and progression-free survival (FPS) on the LNH073B cohort according to 3D ^{18}F-FDG PET/CT image-based features TMTV (cm^3) and Dmax (cm) (A, C), and according to PET MIP image-based features (sTMTV (cm^2) and sDmax (cm)) estimated from AI (B, D).

Tables

TABLE 1. Population characteristics.

Characteristic	REMARC	LNH073B
No. of patients	287	95
Sex		
No. of men	165 (57.5%)	42 (44%)
No. of women	122 (42.5%)	53 (56%)
Median age (y)	68 [64.0-73.0]	46 [33.25-55.0]
Median weight (kg)	72 [63.0-84.2]	68 [58.0-80.0]
Median height (cm)	167.5 [160.0-175.0] (1 case missed)	173 [140.0-193.0]
Ann Arbor stage		
<I	1 (0.4%)	0 (0%)
>=II	286 (99.6%)	95 (100%)
Performance status		
0	115 (40%)	0 (0%)
1	121 (42%)	27 (28.4%)
2	42 (14.6%)	43 (45.3%)
3	2 (0.7%)	20 (21.1%)
4	2 (0.7%)	5 (5.3%)
Missing	5 (1.7%)	NA

Note: data in brackets are interquartile ranges (quartile one to quartile three).

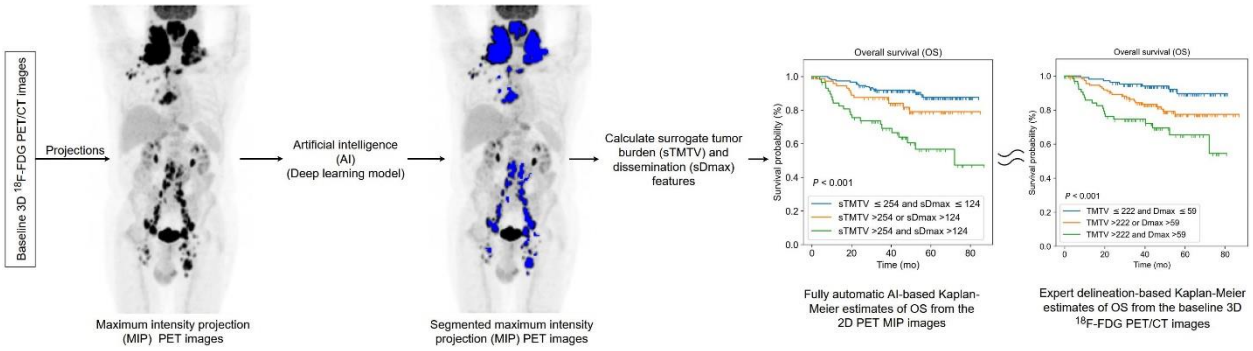
TABLE 2. Statistics for Surrogate TMTV and Surrogate Dmax.

Cohort	sTMTV/sDmax	Mean	SD	Minimum	Q1 (25%)	Median	Q3 (75%)	Maximum
REMARC	sTMTV (cm^2)	252.27	245.75	0.48	77.04	174.24	350.56	1339.36
	sDmax (cm)	100.16	49.89	0.40	66.20	98.0	135.0	225.20
LNH073B	sTMTV (cm^2)	388.12	249.91	63.68	224.48	307.2	450.08	1186.24
	sDmax (cm)	121.82	41.10	43.20	92.00	116.40	145.60	222.40

TABLE 3. Results of the Univariate Analyses for Progression-free Survival (PFS) and Overall Survival (OS) using Time-dependent Area Under the Receiver Operating Characteristics Curve (AUC) analysis and Cox Models (Hazard Ratios (HR)) on the REMARC Data.

			3D ^{18}F-FDG PET/CT estimates		2D PET MIP estimates	
Data	PFS/OS	Metrics	TMTV	Dmax	sTMTV	sDmax
REMARC	PFS	AUC	0.67 (0.60-0.73)	0.65 (0.58-0.72)	0.65 (0.58-0.72)	0.68 (0.62-0.75)
		HR	11.24 (2.10-46.20)	9.0 (2.53-23.63)	11.81 (3.29-31.77)	12.49 (3.42-34.50)
	OS	AUC	0.67 (0.58-0.76)	0.62 (0.53-0.71)	0.67 (0.58-0.76)	0.68 (0.59-0.76)
		HR	16.43 (2.42-77.29)	8.60 (1.47-28.33)	22.14 (4.73-69.06)	22.79 (3.80-79.21)
LNH073B	PFS	AUC	0.62 (0.49-0.75)	0.56 (0.39-0.72)	0.66 (0.53-0.80)	0.58 (0.41-0.74)
		HR	13.79 (0.45-86.80)	32.83 (0.4-220.8)	9.24 (0.95-37.94)	16.79 (0.69-86.41)
	OS	AUC	0.65 (0.46-0.82)	0.51 (0.31-0.72)	0.64 (0.45-0.82)	0.50 (0.29-0.72)
		HR	64.30 (0.74-384.80)	49.21 (0.07-258.3)	14.17 (0.59-67.02)	20.39 (0.08-93.66)

Graphical Abstract



Supplementary material

A. Measurements of Reference TMTV and Dmax

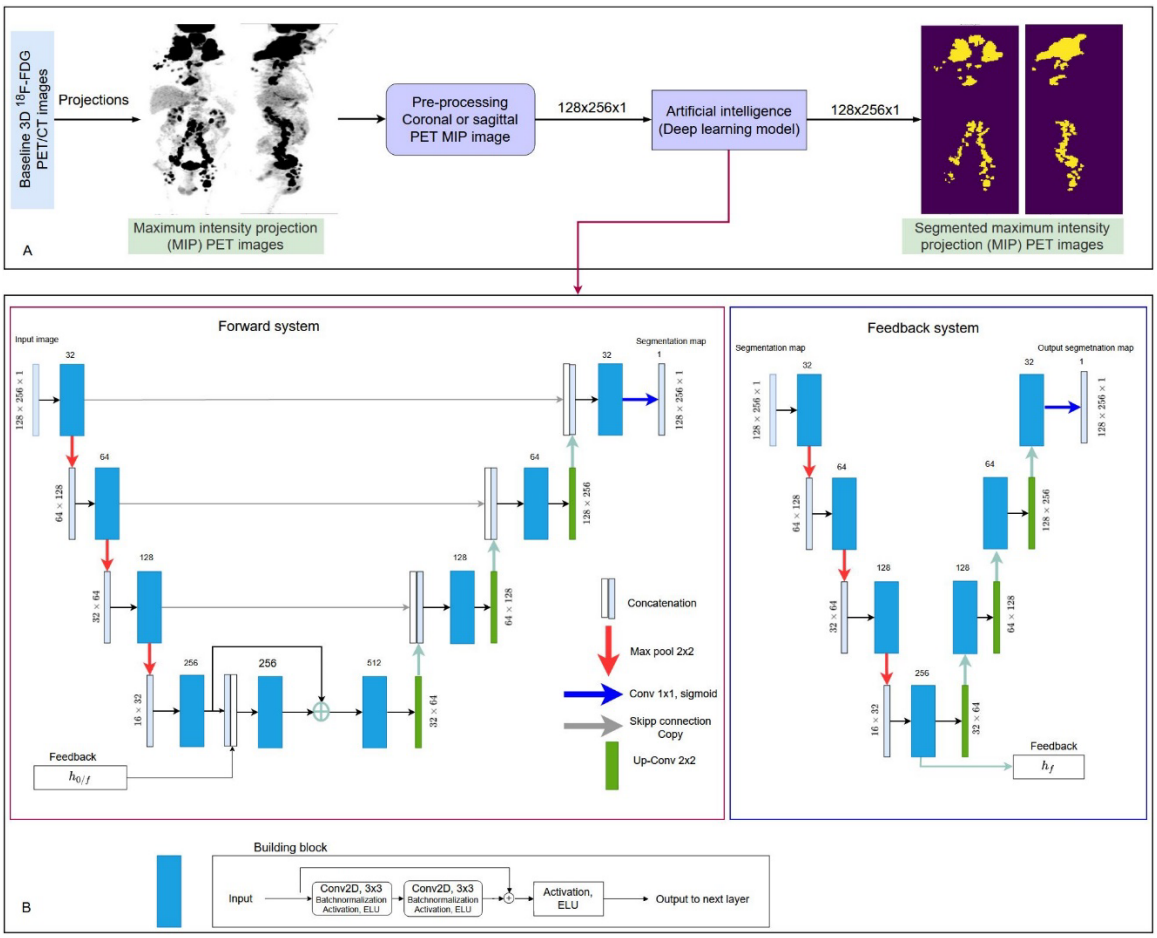
For the REMARC cohort, the lymphoma regions were automatically identified in the 3D PET images as described in (*1*,*2*). A SUVmax 41% threshold segmentation was then applied on these regions, and the results were visually checked by an expert nuclear medicine physician to exclude physiological lesions and to manually add missed lesions whenever needed as described in (*3*).

The LNH073B data were processed by a nuclear medicine physician using the LIFEx software (*4*): hypermetabolic regions were first automatically detected by selecting all voxels with an SUV greater than 2 included in a region greater than 2 mL, and a 41% SUVmax thresholding of the resulting regions was used. Like in the REMARC cohort, the expert removed the regions corresponding to physiological uptakes and added pathological regions missed by the algorithm.

For both cohorts, the physicians were blinded to the patient outcomes. The 3D lymphoma regions validated by experts were used to compute the baseline TMTV and Dmax (based on the centroid of the lymphoma regions) (*5*).

B. Final Network Architecture and Training

The deep learning model was trained from the REMARC data using a five-fold cross-validation technique. It was then tested on another independent cohort, LNH073B. The architecture of the network was inspired by (*6*). The model consists of an encoder and a decoder network with a skipped connection between the two paths and external fully connected network-based feedback. Lymphoma regions are often scattered over the whole body, and information could easily be lost in the successive convolution and pooling operations. To alleviate this scenario, we have used residual CNN as a building block (*7*) in all encoder and decoder components of the deep learning model (Figure 1). It can ease training and facilitate information propagation from input to the output of the network architecture. The input and output dimensions of the network were 128×256×1.



Supplemental Figure 1. Components of the proposed convolutional neural network (CNN). A) Overview of the deep learning model, inputs, and outputs. The coronal or sagittal PET MIP images are provided as independent inputs to the deep learning model. The corresponding segmented regions having the same size as the input image are the output. B) Deep learning model architecture. The building block is the convolutional building block of the deep learning model. Each 2D CNN (Conv2D) with a kernel size of 3x3 was followed by batch normalization and activation function. We have used the exponential linear unit (ELU) activation function, except it was a sigmoid activation function at the output layers. After the convolutional building block in the encoder, we applied a 2x2 max pooling operation with stride 2 for downsampling. Before the convolutional building block, we used a 2x2 up-convolutional layer in the decoder. The deep learning model will be publicly available upon publication [GitHub].

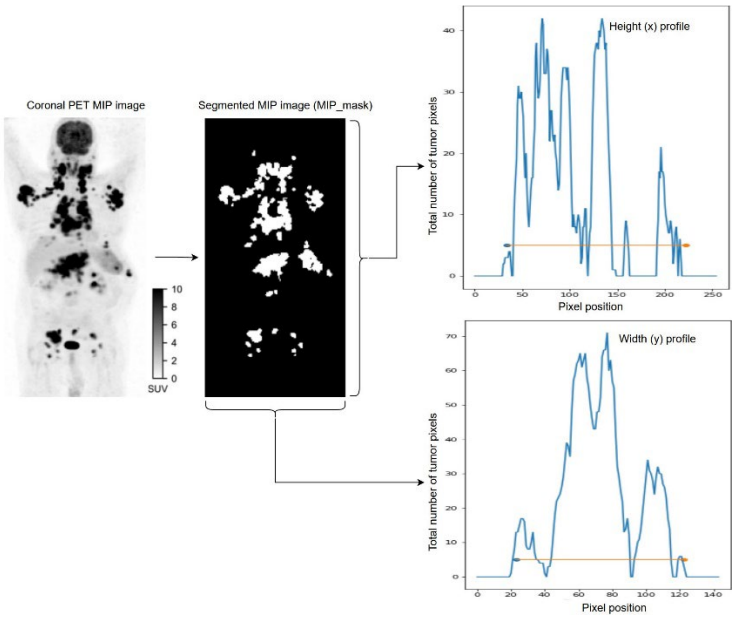
Pre-processing. All available 3D PET images and the corresponding expert-validated 3D lymphoma segmented regions were resized in to 4 x 4 x 4 mm^3 voxel size. The resized 3D images were then padded

or cropped to fit into a 128x128x256. The resized and cropped image were projected into sagittal and coronal views. The input and output image dimensions to the network were 128x256x1.

Training. The model was trained with a batch size of 32 for 1000 epochs and 300 early stop criteria. Different augmentation techniques, including flipping and rotation, were considered and tested but did not improve the results, so we did not use any data augmentation to produce the final model. The deep learning model neural network weights were updated using a stochastic gradient descent algorithm, ADAM optimizer (*8*), with a learning rate of 1e-4. All other parameters were Keras default values. A sigmoid output activation function was used to binarize the image into the lymphoma region and non-lymphoma region. We used the average of the Dice similarity coefficient ($Loss_{Dice}$) and binary cross-entropy ($Loss_{binary\ cross\text{-}entropy}$) as a loss function defined by:

$$loss = 1/2\ (\ Loss_{binary\ cross-entropy} + Loss_{Dice})$$

The model was implemented with Python, Keras API, and Tensorflow backend. The data was processed using the Python 3.8.5 package, including Numpy, Scipy, Pandas, and Matplotlib. We did not apply any post-processing method for the segmentation metrics. To compute the surrogate biomarkers from the AI-based segmented images, regions with less than 4.8 cm^2 were removed. The deep learning model will be publicly available upon publication at [GitHub].



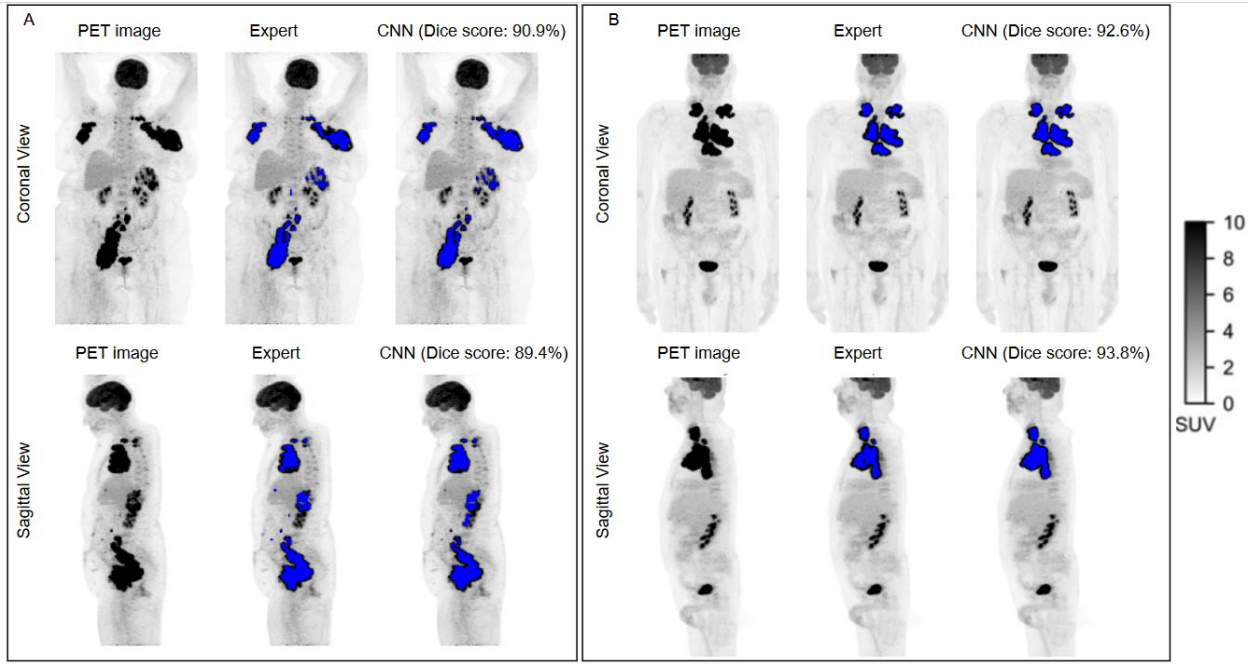
Supplemental Figure 2. Illustration of the calculation of the tumor dissemination feature. For the given PET MIP image, we created two profiles corresponding to the sum of the signal in the x and y directions, respectively. The horizontal line shows the distances between the 2% percentiles and the 98% percentiles. It was the same for the sagittal PET MIP image. Pixel positions with zero total number of tumor pixels (often at the beginning and end of the pixel positions) are not considered for the percentile calculation.

C. Statistical Analysis Details

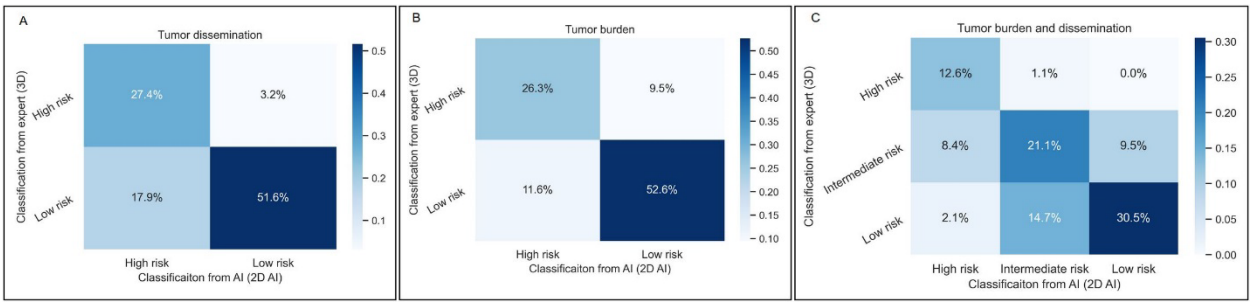
Univariate Analysis. For all biomarkers, we calculated a time-dependent area under the receiver operating characteristics curve (AUC) (*9*). Bootstrap resampling analysis was performed to associate confidence intervals to the Cox model hazard ratio and the time-dependent AUC. The bootstrapping involved 10,000 random samplings of the data with replacement. All statistical comparisons, except the Kaplan-Meier analysis, were made without discretizing the continuous values.

Multivariate Analysis. We estimated the survival functions using Kaplar-Meier estimates. For each PET-derived feature, we selected the optimal cut-off values for PFS and OS at the values that yielded the smallest P-value in the log-rank test between categories of a given study population. The cut-off values were constrained to be between the interquartile ranges of the TMTV or Dmax values. This procedure was the same for all measurements, namely for the 3D ^{18}F-FDG PET-based biomarkers (TMTV and Dmax) and

PET MIP-based biomarkers from the deep learning method (sTMTV and sDmax). A receiver-operating-characteristics (ROC) analysis was also used to define the optimal cut-off values that predict the occurrence of an event (progression-free survival or overall survival) by maximizing the sensitivity plus specificity minus one (i.e., sensitivity + specificity -1). It yielded nearly the same results as calculating the cut-off values using the log-rank test approach. For the TMTV, we obtained a cut-off value of 222 cm^3, which is close to the published values of 220 cm^3 (*1*). For uniformity of the comparison of the 2D and 3D PET features, we followed the same procedures for all features to compute the cut-off values.



Supplemental Figure 3. ^{18}F-FDG PET MIP images and segmentation results (blue color overlapped over the PET MIP images) by experts (MIP_masks) and by the CNN for four patients: (A) from the REMARC cohort, and (B) from the LNH073B cohort.



Supplemental Figure 4. Confusion matrices for classification of patients using PET features derived from using the expert-delineated 3D ^{18}F-FDG PET images (3D-expert) and from using the 2D PET MIP images using CNN (2D-AI) on LNH073B cohort. A) Two-risk-group classification using Dmax and sDmax, B) two-risk-group classification using TMTV and sTMTV, and C) three-risk-group classification using TMTV and Dmax (3D-expert), and sTMTV and sDmax (CNN).

References

1. Vercellino L, Cottereau AS, Casasnovas O, et al. High total metabolic tumor volume at baseline predicts survival independent of response to therapy. *Blood*. 2020;135:1396-1405.
2. Capobianco N, Meignan M, Cottereau A-S, et al. Deep-learning 18 F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-Cell lymphoma. *J Nucl Med*. 2021;62:30-36.
3. Cottereau A-S, Nioche C, Dirand A-S, et al. 18F-FDG PET dissemination features in diffuse large B-cell lymphoma are predictive of outcome. *J Nucl Med*. 2020;61:40-45.
4. Nioche C, Orlhac F, Boughdad S, et al. Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res*. 2018;78:4786-4789.
5. Cottereau A-S, Meignan M, Nioche C, et al. Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†. *Ann Oncol*. 2021;32:404-411.
6. Girum KB, Crehange G, Lalande A. Learning with context feedback loop for robust medical image segmentation. *IEEE Trans Med Imaging*. 2021;40:1542-1554.
7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:770-778.
8. Kingma DP, Ba J. Adam: A method for stochastic optimization. *3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc*. December 2014:1-15.
9. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56:337-344.