Revista Brasileira de Física Médica (2023) 17:680

Prediction of Treatment Failure in Hodgkin Lymphoma: A Machine Learning Radiomic Approach in Baseline 18F-FDG PET/CT

Predição de Falha no Tratamento de Linfoma Hodgkin: Uma Abordagem de Aprendizado de Máquinas em 18F-FDG PET/CT

Marcos A. D. Machado[1,2], Cleiton C. Queiroz [1,3,4], Jean L. P. Pita[4], Lucas O. Vieira[1], Bruno C. F. R. Santana[1], Mauro Namías[5], Vinicius O. Menezes[1,6], Marco A. Salvino[2,7], Simone C. S. Brandão[6], Eduardo M. Netto[2]

[1]Nuclear Medicine Department, São Rafael Hospital, Salvador, Brazil;
²Complexo Hospitalar Universitário Prof. Edgard Santos da Universidade Federal da Bahia/Ebserh, Salvador, Brazil;
[3]Hospital Universitário Prof. Alberto Antunes da Universidade Federal de Alagoas/Ebserh, Salvador, Brazil;
[4]Department of Computer Science, Universidade Federal de Bahia, Brazil;
[5]Fundación Centro Diagnóstico Nuclear, Buenos Aires, Argentina;
[6]Hospital das Clínicas da Universidade Federal de Pernambuco/Ebserh, Recife, Brazil;
[7]Hemathology Department, São Rafael Hospital, Salvador, Brazil

Resumo
Objetivo: A avaliação do estadiamento e prognostico do Linfoma Hodgkin usando métodos de aprendizado de máquinas ainda é pouco explorado. Este estudo explora características radiômicas e propõe um novo modelo de aprendizado de máquinas para predizer falha no tratamento a partir de imagens PET. Métodos: Treze características radiômicas foram extraídas de imagens PET de 51 indivíduos com Linfoma Hodgkin com resposta ao tratamento, e 6 indivíduos com falha no tratamento de acordo com o critério de Lugano. Foram feitas análises univariadas das áreas abaixo da curva ROC (AUC) para a seleção de características radiômicas, e a correlação de Pearson foi usada para avaliar a redundância entre as características radiômicas. Um algoritmo de aprendizado de máquinas foi desenvolvido usando 6 grupos de indivíduos, cada grupo contendo 1 indivíduo refratário ao tratamento. Os 6 grupos foram combinados em um esquema de validação cruzada com 15 camadas usando amostragem aleatória com reposição dos indivíduos com resposta ao tratamento. Cada camada contendo 4 grupos para construção do modelo e 2 grupos para validação. O desempenho do algoritmo foi verificado a partir dos resultados das 15 combinações de cada uma das 2 camadas de validação, usando dois métodos (M1 e M2) para classificação do tumor, tanto no nível da lesão quanto no nível do paciente. Resultados: As características radiômicas *zone percentage (ZP), high intensity large area emphasis (HILAE), entropy, complexity* e *standardized uptake value of maximum pixel (SUVmax)* foram preditores independentes de falha no tratamento. O algoritmo de aprendizado de máquina desempenhou com AUC=0,86 (M1) e AUC=0,96 (M2) para classificar “tumores individuais”, e no “nível paciente” desempenhou com sensibilidade = 80,0%/ especificidade = 88,3% (M1) e sensibilidade = 100%/ especificidade = 100% (M2). O algoritmo desempenhou melhor que o método de estadiamento *Ann Arbor,* que alcançou, respectivamente, 83,3% e 31,7% para sensibilidade e especificidade. Conclusões: A avaliação de imagens basais de 18F-FDG PET de pacientes com Linfoma Hodgkin usando aprendizado de máquinas obteve melhor desempenho em relação ao estadiamento Ann Arbor e em relação aos marcadores MTV e SUVmax, possibilitando melhorar a classificação de pacientes com maior risco de falha no tratamento.

Palavras-chave: imagem molecular, aprendizado de máquinas, linfoma Hodgkin, medicina de precisão.

Abstract
Purpose: Hodgkin Lymphoma staging and prognostic evaluation through radiomic and machine learning (ML) methods have been little explored. This study explores radiomic features and proposes a new ML radiomic model to predict treatment failure in Hodgkin lymphoma patients from PET images. Methods: 13 radiomic features were extracted from PET images of 51 subjects with Hodgkin's Lymphoma with response to therapy and 6 subjects with treatment failure as classified by the Lugano criteria for lymphoma response. Univariate analysis was performed for feature selection employing the areas under ROC curves (AUC), and Pearson correlation was assessed to reduce redundancy in the feature space. An ML algorithm was developed and trained-validated using 6 groups of subjects, each group containing 1 treatment-refractory individual. The 6 groups were combined in a 15-fold cross-validation using random sampling with the replacement of non-progressive disease subjects at each fold. Each fold included 4 groups for model building and 2 groups for model validation. The ML performance was assessed through 15 combinations of two-fold validation samples, deriving two ML methods (M1 and M2) for tumor classification at the tumor and patient levels. Results: The features zone percentage (ZP), high-intensity large area emphasis (HILAE), entropy, complexity, and standardized uptake value of maximum pixel (SUVmax) were independent predictors of treatment failure. The ML algorithm performed with AUC=0.86 (M1) and AUC=0.96 (M2) to classify "individual tumors" and at "patient level" performed with sensitivity = 80.0% /specificity = 88.3% (M1) and sensitivity = 100% /specificity = 100% (M2). The ML outperformed the Ann Arbor staging that achieved, respectively, 83.3% and 31.7% for sensitivity and specificity. Conclusions: Evaluation of baseline 18F-FDG PET scans of individuals with Hodgkin's lymphoma using an ML radiomic model performed better than Ann Arbor staging and MTV and SUVmax, allowing improved classification of patients at higher risk of treatment failure.

Keywords: *molecular imaging, machine learning, Hodgkin lymphoma, precision medicine.*

https://doi.org/10.29384/rbfm.2023.v17.19849001680

1. Introduction

Positron emission tomography (PET) and computed tomography (CT) hybrid imaging (PET/CT) using ^{18}F–fluorodeoxyglucose (^{18}F-FDG) holds a major place in staging Hodgkin Lymphoma (HL) with the ability to detect the extent of HL and bone marrow involvement allowing a non-invasive procedure for the staging of nodal and extra-lymphatic disease. Furthermore, ^{18}F-FDG PET/CT also introduced a significant improvement in response evaluation when compared to conventional imaging[1]. In addition to monitoring changes during treatment, recent studies have shown that baseline PET has the potential to predict response to therapy by measuring the total metabolic tumor volume (TMTV) biomarker [2,3].

Classical measurements of tumor uptake (SUV and TMTV) ignore the intratumoral ^{18}F-FDG spatial distribution (i.e., tumor heterogeneity). The rapidly emerging field of "radiomics" computes quantitative image features that characterize this intratumoral heterogeneity or other tumor phenotypes, leading to a promising field in the era of precision medicine. There is growing evidence that radiomics could add useful information to baseline PET, CT, or MR to predict outcomes in some types of cancer [4]. Bouallègue *et al.* investigated the feasibility of textural and morphological features in predicting early metabolic response assessed from baseline ^{18}F-FDG PET on a heterogeneous cohort of Hodgkin and non-Hodgkin lymphoma patients [5]. Only a few studies have demonstrated that machine learning (ML) from baseline PET radiomics in HL can predict treatment outcomes [6,7], suggesting that radiomics may enhance the predictive capability of conventional metrics. On the other hand, Lartizien and colleagues demonstrated the use of radiomics ^{18}F-FDG PET/CT in HL patients to differentiate cancer lesions from hypermetabolic inflammatory foci [8].

In this work, we explored radiomic features extracted from baseline ^{18}F-FDG PET scans of HL patients and developed ML classifiers to predict treatment failure.

2. Methods

2.1. Scanner, imaging, and patient selection

^{18}F-FDG PET/CT imaging was performed on a PET Siemens Biograph TruePoint TrueV (Knoxville, TN, USA) combined with a 16-slice helical CT scanner (Emotion 16; Siemens). PET images were corrected for random coincidences, normalization, dead time losses, scatter, and attenuation.

Interim PET (iPET) of HL patients is recommended most often after two cycles to assess early response to therapy. The purpose is to confirm the effectiveness of treatment, using iPET to tailor treatment according to individual response[1]. As the time of iPET is not mandatory, clinicians in our institution refer patients for response evaluation either during treatment (iPET) or after the treatment is over (post-therapy, pt-PET). Hereafter, we will consider both as pt-PET. Then HL ^{18}F-FDG PET/CT exams that performed both baseline and pt-PET within 12 months from the beginning of treatment were selected (mean: 4.5, range: 1.4-11.8). Exclusion criteria were hyperglycemia at the time of tracer administration or a delay exceeding 90 min between ^{18}F-FDG injection and image acquisition. A total of 57 subjects were selected. An intravenous FDG injection and acquisition duration was performed according to the European Association of Nuclear Medicine (EANM) guidelines[9] and applying the method previously reported by Menezes et al. for noise standardization[10,11]. Subjects fasted for 6 hours prior to the ^{18}F-FDG injection. The median injection-to-scan time was 66 minutes (standard deviation 10 minutes). Images were reconstructed on a 168x168 matrix size (4.07 x 4.07 mm² voxels) with 3 mm slice thicknesses and using OSEM3D (ordered-subsets expectation maximization) algorithm with 3 iterations, 21 subsets, and a 5 mm gaussian filter. These parameters provide quantitatively harmonized images according to EANM.

The study population was selected based on the International Classification of Diseases, Tenth Revision (ICD-10) (C81.X) for HL. The pt-PET was evaluated according to Lugano classification lymphoma response criteria, a score derived from the Deauville 5-point scale (5PS) which scores 1 to 3 are considered Complete Metabolic Response (CMR), scores 4 and 5 are considered either No-Metabolic-Response (NMR), Partial Response (PR), or Progressive Disease (PD) depending on the evolution compared to baseline evaluation (no change from baseline, reduced uptake compared to baseline, and increased uptake or new lesion, respectively). This classification results in standardized reports with good prognosis discrimination and inter-observer reproducibility [1]. We also conducted a review of each patient's electronic medical record to determine their initial status and outcome at pt-PET.

The study was approved by the Hospital São Rafael ethical board (CAAE: 79258017.6.0000.0048), waived the informed consent, and all information was anonymized. The images were acquired according to the clinical protocol of Hospital São Rafael for tumor PET imaging.

2.2. Radiomic feature extraction

Individual lesions from baseline PET scans were segmented with the Beth Israel PET/CT plugin for FIJI (ImageJ, Bethesda, MD, USA) [2,12] using a 41% threshold of maximum SUV (SUV_{max}). Segmented VOIs were analyzed with PyRadiomics, an open-source platform available at www.radiomics.io that enables the extraction of a large panel of radiomic features [13]. We resampled the matrix grid to cubic voxels of 4x4x4 mm³ and used intensity discretization with a fixed bin size of 0.25 g/ml to avoid feature instability[14]. The radiomic feature classes and corresponding features are presented in Table 1 as defined by Griethuysen *et al.* [13]

Table 1: Radiomic features.

Class	Features
First-order statistics	SUV_{max}, kurtosis
Shape descriptors	Metabolic tumor volume (MTV), sphericity, surface volume ratio (SVR)
Gray Level Co-occurrence Matrix (GLCM)	Entropy (E), difference average (D)
Gray Level Size Zone Matrix (GLSZM)	Zone percentage (ZP), high-intensity large area emphasis (HILAE)
Neighboring Gray Tone Difference Matrix (NGTDM)	Coarseness, complexity

Thirteen features from five classes were selected from previous studies [5,6,8,14–17]. Entropy and difference average from GLCM were calculated according to 2 different methods as described in Hatt *et al.* (2015), which will be denoted as E13 and D13, where E is entropy and D is difference average, whereas the parameters from the second method will be denoted as E1 and D1. Lesions with an MTV < 5 cm³ were excluded from the analysis because of insufficient voxels to compute textural indexes.

2.3. Tumor classification

In pt-PET, patients diagnosed as PD represent the group at higher risk of treatment failure and death [6,18,19]. We, therefore, classified the baseline VOIs based on the clinical assessment of pt-PET: CMR, NMR, and PR were annotated as the responder group; and PD was annotated as the non-responder group.

2.4. Statistical analysis

Tumor level: Analysis for the association between features and response to therapy status was performed using the area under the ROC curve (AUC) computation. Features with p-value < 0.05 were selected. Pearson correlation was used to reduce the space by further selecting the features which r < 0.90. The population was randomly divided into 6 groups, each group containing one subject with progressive disease. All tumors from four groups were used to train a Random Forest-based AdaBoost enhanced machine learning (ML) model to discriminate tumors into two classes (progressive disease vs. non-progressive disease), and the two remaining groups of tumors were used for model validation. To estimate the classifier performance, we combined the 6 groups into 15-fold cross-validation, each fold using 4 groups for model building and 2 folds for model validation, and using random sampling with replacement of non-progressive disease subjects at each fold (supplementary material). The Python library Scikit-learn was used to implement these methods, classifying tumors through an ML score ranging from 0 to 1, with 0 meaning non-progressive disease and 1 meaning progressive disease[20].

Subject level: Prediction to discriminate responders and non-responders at the subject level was assessed through the tumor with the maximum ML score within each patient, assessing two methods: **M1**) direct computation of sensitivity and specificity at every independent fold: subjects which scored higher than ML score threshold *S* were considered as progressive disease. The sensitivities and specificities were averaged among all folds; and **M2**) averaging the scores from the 15 folds for each tumor (S_m), thus analyzing all 57 subjects together. Subjects who scored higher than S_m were considered as having progressive disease. The sensitivity and specificity were computed for the whole dataset.

S and S_m were determined through Youden's index. The statistical power 1-β of each method was assessed with a significance level α = 5%. Fisher's exact test was performed to verify differences between groups for categorical variables, and Wilcoxon Mann–Whitney test for continuous data. Also, confidence intervals of sensitivities and specificities were estimated using binomial proportion Wilson's score method. Prognostication was compared with Ann Arbor staging for the initial stage (I or II) and advanced stage (III or IV). SPSS 21.0 for Windows or OpenEpi (Emory University) was used for statistical analysis.

3. Results

Population characteristics are presented in Table 2.

3.1. Tumor level

Feature selection (Table 3) and reduction (Figure 1) revealed five radiomic predictors of treatment failure: SUVmax, E1, ZP, HILAE, and Complexity. D1 and D13 were also predictors, but due to their high correlation with SUVmax, they can be represented by SUVmax. ML algorithm thus constructed with these predictors and validated in every 15-independent fold accomplished superior discrimination than individual features. Detailed performance at every 15-validation fold is presented in supplementary material, where the mean AUC of M1 at the lesion level, AUC_m = 0.86. For improved readability and interpretation of results, the mean ROC curve for M1 (ROC_m) was simulated by iteratively misspecifying the true values with a random gaussian function. The procedure was stopped when the simulated AUC_S = AUC_m = 0.86 [21]. Figure 2 demonstrates the performance of univariate features and ML classifiers at the tumor level M1 (AUC_s = 0.86, 95% CI: 0.77-0.94) and M2 (AUC=0.96, 95% CI: 0.93-1.00).

Table 2: Population characteristics.

Characteristics	Responders (n=51)	Non-responders (n=6)	p-value
Median age (range)	28 (4-82)	18 (7-29)	0.04
Male	53%	50%	1.00
Ann Arbor			0.66
stage I/ stage II	2% /29%	0% /17%	
stage III/ stage IV	26% /43%	33% /50%	
B symptoms	88% [a]	67% [b]	0.40
Bulk (>10 cm)	59%	50%	0.69
Extra nodal disease	65%	50%	1.00
ABVD or ABVD-like chemotherapy	100%	100%	1.00
Median time (month) §	4.7 (1.4-11.3)	4.5 (1.8-11.8)	0.94

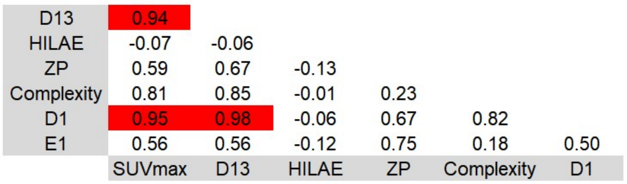
Valid values computed: 10 subjects with missing values [a], 3 subjects with missing values [b].

§ Period between baseline and pt-PET

Table 3: Feature statistics at the lesion level

Features	Responder (n=289)		Non-responder (n=31)		p-value
	IQR	Median	IQR	Median	
MTV (cm³)	8.8-31.7	14.7	7.6-38.7	15.7	ns
SUV_{max}	6.6-12.1	8.9	4.6-8.5	7.1	**
Kurtosis	2.4-3.5	2.8	2.5-3.2	2.8	ns
D13	3.1-6.0	4.5	1.8-4.0	3.4	**
D1	3.0-6.2	4.4	1.7-3.9	3.4	**
E13	6.3-7.8	7.0	5.5-7.8	6.4	ns
E1	7.1-8.9	7.9	5.7-8.0	7.6	***
ZP	0.36-0.63	0.53	0.24-0.47	0.41	**
HILAE	3.8-19.0	6.7	10.5-116.3	17.0	**
Coarseness	0.01-0.04	0.03	0.01-0.05	0.03	ns
Complexity	286-1844	720	52-611	349	**
Sphericity	0.53-0.70	0.63	0.54-0.75	0.65	ns
SVR	0.26-0.38	0.32	0.25-0.39	0.32	ns

IQR: interquartile range; MTV: metabolic tumor volume; SUV_{max}: standardized uptake value of maximum pixel; D13: difference average using 13 matrices; D1: difference average using 1 matrix; E13: entropy using 13 matrices; E1: entropy using one matrix; ZP: zone percentage; HILAE: high-intensity large area emphasis; SVR: surface volume ratio; p-value of AUC: ** $p \leq 0.001$, *** $0.001 < p < 0.05$, ns: $p \geq 0.05$.



D13	0.94					
HILAE	-0.07	-0.06				
ZP	0.59	0.67	-0.13			
Complexity	0.81	0.85	-0.01	0.23		
D1	0.95	0.98	-0.06	0.67	0.82	
E1	0.56	0.56	-0.12	0.75	0.18	0.50
	SUVmax	D13	HILAE	ZP	Complexity	D1

Figure 1: Correlation matrix of best features predictors. Features with Pearson correlation r > 0.9 are highlighted.

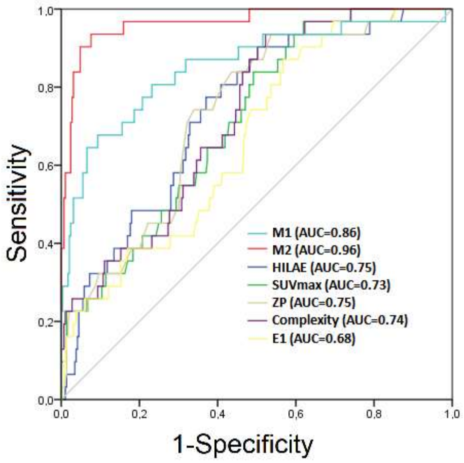


Figure 2: ROC curves of univariate features and machine learning classifier.

3.2. Subject level

Upon translating the classification to the subject level, the sensitivity and specificity were, respectively, 80.0% (CI = 43.7-97.0%) and 88.3% (76.6-94.5%) for M1 (S = 0.55) and 100% (70.0-100.0%) and 100% (93.0-100%) for M2 (S_m = 0.58) (Table 4). Figure 3 illustrates pretreatment ML scores among groups of pt-PET status at the tumor level (A) and subject level (B) for M2. Both methods, M1 and M2, were strong predictors of treatment failure, which the estimated statistical powers 1-β were 94% and 100%, respectively.

Ann Arbor staging had 83.3% (CI = 43.7-97.0%) and 31.7% (CI = 20.3-45.0%) sensitivity and specificity, respectively. Contrary to the ML scores, it was not possible to detect a difference between groups for Ann Arbor staging (p = 0.66).

An example of radiomic characteristics of two lesions in Figure 3-B was ML score = 0.77 (SUVmax = 3.7, ZP = 0.17, Complexity = 11, HILAE = 150 and E1 = 4.5) and ML score = 0.30 (SUVmax = 7.8, ZP = 0.54, Complexity = 744, HILAE = 4.7 and E1 = 8.4).

Table 4: Classifier performance of machine learning radiomics and Ann Arbor at subject level.

	Sensitivity (95% CI)	**Specificity** (95% CI)
M1	80.0% (43.7-97.0%)	88.3% (76.6-94.5%)
M2	100% (70.0-100%)	100% (93.0-100%)
Ann Arbor	83.3% (43.7-97.0%)	31.7% (20.3-45.0%)

CI: confidence interval

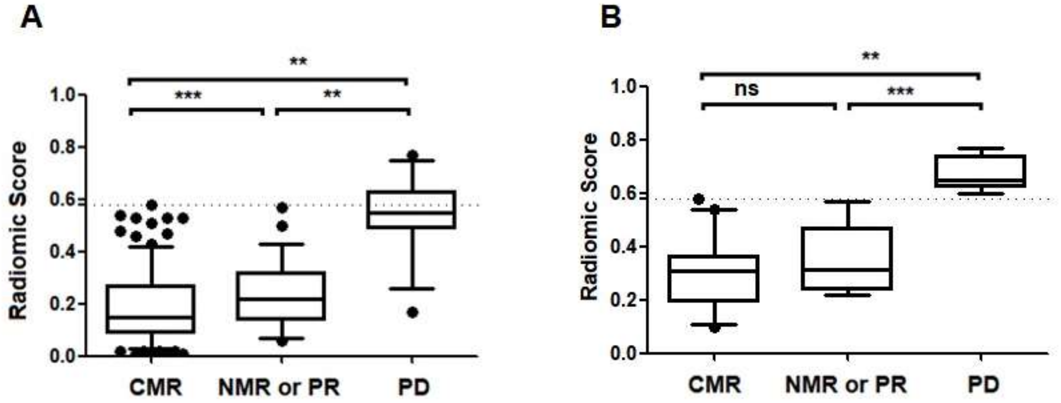


Figure 3: ML score of M2 versus pt-PET status. A: ML scores for individual tumors. B: maximum ML tumor score for each subject. Boxplots represent 5-95% distribution. Dashed lines represent the optimal score threshold of 0.58. CMR: complete metabolic response, NMR or PR: no metabolic or partial response, and PD: progressive disease. ** $p < 0.001$. *** $p < 0.05$. ns: $p \geq 0.05$.

4. Discussion

We evaluated new risk factors of treatment failure and predictive models through baseline ^{18}F-FDG PET/CT of HL by characterizing tumor ^{18}F-FDG distribution. We conducted a careful methodological design and data quality control to avoid possible pitfalls in the radiomic workflow and analysis and assure the reliability and repeatability of the results [14-17].

Recent studies of morphological biomarker TMTV in baseline ^{18}F-FDG PET/CT demonstrated that a higher TMTV indicates the worst prognosis [2,3]. Hatt et al. (2015) studied associations between textural features and MTV in various tumors, demonstrating the added contribution and dependence of textural features on MTV [15]. Here, we evaluated tumor texture as an indicator of disease progression and excluded MTV in our ML model since the study protocol anticipated the exclusion of features with $p \geq 0.05$ for the AUC. Current evidence using radiomics with HL ^{18}F-FDG PET/CT suggests that robust statistical approaches based on ML would potentially improve predictive models for HL staging[6,7]. Thus, we conducted two ML radiomic classification methods with good performance using a different approach by testing all six "progressive disease" patients through 15 independent validation folds.

Radiomic features and tumors have a complex relationship. The exact relationship between radiomic features and underlying tumor biology can be established only on carefully designed prospective studies. However, we can assume that a combination of features may be associated with tumor cellularity, vascularization, perfusion, proliferation, aggressiveness, hypoxia, angiogenesis, and necrosis, factors related to more aggressive behavior, poorer response to treatment, and worse prognosis[24,25]. We demonstrated that five features (out of 13 pre-selected features) are informative regarding differences in ^{18}F-FDG uptake heterogeneity in HL ^{18}F-FDG PET/CT. From table 3, texture features HILAE, ZP, E1, and Complexity suggest that more homogeneous tumors are less likely to respond to therapy than more heterogeneous ones. In this study, M1 was assessed through every 15 independent folds, while M2 was not validated in an independent dataset. Then, M2 may have been positively biased and lacks internal validation [17]. Additionally, feature selection and reduction should ideally be performed using only the model development dataset [17], but it was not possible in this study due to our cross-validation approach using all six "progressive disease" patients. Indeed, the results of M1 and M2 are encouraging to validate or refute this pilot study in an external cohort.

Our study has other limitations. First, we used the Lugano Criteria as the gold standard to define treatment failure, which has 73% sensitivity and 94% specificity for iPET [9,10]. Since end-of-treatment PET is more meaningful than iPET, tumor classification bias probably occurred. However, the high specificity of our model could diminish this impact while translating the diagnostic information to the subject level. Second, the study involved only one center and PET system, lacking external validation. However, we used reconstruction settings in compliance with the EANM standards[9,] and discretization with a fixed bin size is known to provide more robust results[14]. Despite the selected features being reported robust in test–retest reproducibility[16], our results are valid only for the image reconstruction, segmentation (41% threshold of SUVmax increases the weight of regions with high uptake), and quantification settings presented herein, since these settings significantly affect the values of radiomic features [27]. In addition, patient characteristics were not homogenous (table 2), with non-responders being younger than responders (p = 0.04), possibly introducing patient selection bias. Last, we used a small sample size, despite being sufficiently large to detect differences in radiomic profile among subject groups and powerful enough to model the complex relationship among features.

Our efforts are in line with a global demand to make radiomics and molecular imaging feasible and translational to clinical practice, as methodological procedures, image noise standardization, and feature harmonization is becoming more popular [28,29].

5. Conclusions

Our study indicates that radiomic features from baseline ^{18}F-FDG PET scans of HL are associated with treatment failure by Lugano Criteria and with poorer prognosis. The proposed method outperformed conventional PET metrics (SUVmax and MTV) and the current Ann Arbor staging. This association enabled a radiomic ML classification that offers promise for personalized medicine, thus stratifying patients to conduct the best treatment strategies.

Acknowledgments

We thank Salim Kanoun for helpful instructions on managing the PET/CT Viewer Beth Israel plugin for FIJI and integrations with Pyradiomics.

Funding

This work was supported, in part, by the Federal University of Bahia (Permanecer project and Programa de Pós-Graduação em Medicina e Saúde, Faculdade de Medicina da Bahia, Salvador, Bahia, Brazil) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES).

References

1. Kanoun S, Rossi C, Casasnovas O. [18F]FDG-PET/CT in hodgkin lymphoma: Current usefulness and perspectives. Cancers (Basel). 2018;10(5):1-11. doi:10.3390/cancers10050145
2. Mettler J. Metabolic Tumour Volume for Response Prediction in Advanced-Stage Hodgkin Lymphoma. J Med Imaging. 2018;i:1-25. doi:10.2967/jnumed.118.210047
3. Cottereau AS, Versari A, Loft A, et al. Prognostic value of baseline metabolic tumor volume in early-stage Hodgkin lymphoma in the standard arm of the H10 trial. Blood. 2018;131(13):1456-1463. doi:10.1182/blood-2017-07-795476
4. Guiot, Julien, et al. "A review in radiomics: Making personalized medicine a reality via routine imaging." Medicinal Research Reviews 42.1 (2022): 426-440.
5. Bouallègue FB, Tabaa YA, Kafrouni M, Cartron G, Vauchot F, Mariano-Goulart D. Association between textural and morphological tumor indices on baseline PET-CT and early metabolic response on interim PET-CT in bulky malignant lymphomas. Med Phys. 2017;44(9):4608-4619. doi:10.1002/mp.12349
6. Milgrom SA, Elhalawani H, Lee J, et al. A PET Radiomics Model to Predict Refractory Mediastinal Hodgkin Lymphoma. 2019;(December 2018):1-8. doi:10.1038/s41598-018-37197-z
7. Frood, R., Clark, M., Burton, C. et al. Utility of pretreatment FDG PET/CT–derived machine learning models for outcome prediction in classical Hodgkin lymphoma. Eur Radiol 32, 7237–7247 (2022). https://doi.org/10.1007/s00330-022-09039-0
8. 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 Heal Informatics. 2014;18(3):946-955. doi:10.1109/JBHI.2013.2283658
9. Boellaard R, Delgado-bolton R, Oyen WJG, et al. FDG PET / CT: EANM procedure guidelines for tumour imaging: version 2. 0. 2015:328-354. doi:10.1007/s00259-014-2961-x
10. Menezes V, Machado M, Queiroz C, et al. How does Lean Six Sigma method improve healthcare practice in nuclear medicine departments? A successful case of dedicated software applications in oncological PET/CT. J Nucl Med. 2018;59(supplement 1):1013-1013. http://jnm.snmjournals.org/cgi/content/short/59/supplement_1/1013. Accessed April 16, 2019.
11. Menezes VO, Machado MAD, Queiroz CC, et al. Optimization of oncological 18 F-FDG PET/CT imaging based on a multiparameter analysis. Med Phys. 2016;43(2):930-938. doi:10.1118/1.4940354
12. Kanoun S, Tal I, Berriolo-riedinger A, Rossi C. Influence of Software Tool and Methodological Aspects of Total Metabolic Tumor Volume Calculation on Baseline [18F] FDG PET to Predict Survival in Hodgkin Lymphoma. 2015;C:1-15. doi:10.1371/journal.pone.0140830
13. Griethuysen J, Fedorov A, Parmar C, Hosny A, Narayan V. Computational Radiomics System to Decode the Radiographic Phenotype. HHS Public Access. Cancer Res. 2018;77(21):1-8. doi:10.1158/0008-5472.CAN-17-0339.
14. Pfaehler E, Jong JR De, Boellaard R. Repeatability of 18 F-FDG PET radiomic features : A phantom study to explore sensitivity to image reconstruction settings, noise , and delineation method. 2018;0(0):1-14. doi:10.1002/mp.13322
15. Hatt M, Majdoub M, Vallieres M, et al. 18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort. J Nucl Med. 2015;56(1):38-44. doi:10.2967/jnumed.114.144055
16. Galavis P, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. NIH Public Access. Acta Oncol. 2010;49(7):1012-1016. doi:10.3109/0284186X.2010.498437
17. Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging. 2019. doi:10.1007/s00259-019-04391-8
18. Barrington SF, Mikhaeel NG, Kostakoglu L, et al. Role of imaging in the staging and response assessment of lymphoma: Consensus of the international conference on malignant lymphomas imaging working group. J Clin Oncol. 2014;32(27):3048-3058. doi:10.1200/JCO.2013.53.5229
19. Biggi A, Gallamini A, Chauvie S, et al. International Validation Study for Interim PET in ABVD-Treated, Advanced-Stage Hodgkin Lymphoma: Interpretation Criteria and Concordance Rate Among Reviewers. J Nucl Med. 2013;54(5):683-690. doi:10.2967/jnumed.112.110890
20. Pedregosa F, Weiss R, Brucher M. Scikit-learn : Machine Learning in Python. 2011;12:2825-2830.
21. Kundu S, Kers JG, Janssens ACJW. Constructing Hypothetical Risk Data from the Area under the ROC Curve : Modelling Distributions of Polygenic Risk. 2016:1-11. doi:10.1371/journal.pone.0152359
22. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? Eur J Nucl Med Mol Imaging. 2017;44(1):151-165. doi:10.1007/s00259-016-3427-0
23. Vallieres M, Zwanenburg A, Badic B, Cheze-Le Rest C, Visvikis D, Hatt M. Responsible Radiomics Research for Faster Clinical Translation. J Nucl Med. 2017:jnumed.117.200501. doi:10.2967/jnumed.117.200501
24. Cook GJR, Yip C, Siddique M, et al. Are Pretreatment 18F-FDG PET Tumor Textural Features in Non-Small Cell Lung Cancer Associated with Response and Survival After Chemoradiotherapy? J Nucl Med. 2013;54(1):19-26. doi:10.2967/jnumed.112.107375
25. Tixier F, Le Rest CC, Hatt M, et al. Intratumor Heterogeneity Characterized by Textural Features on Baseline 18F-FDG PET Images Predicts Response to Concomitant Radiochemotherapy in Esophageal Cancer. J Nucl Med. 2011;52(3):369-378. doi:10.2967/jnumed.110.082404
26. Coroller TP, Agrawal V, Huynh E, et al. Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. J Thorac Oncol. 2017;12(3):467-476. doi:10.1016/j.jtho.2016.11.2226
27. Machado, MAD, Pita, JLP., Netto, EM. (2021). PET Radiomic Features Variability: A Phantom Study on the Influence of Reconstruction and Discretization Method. Revista Brasileira De Física Médica, 15, 598. doi: https://doi.org/10.29384/rbfm.2021.v15.19849001598.

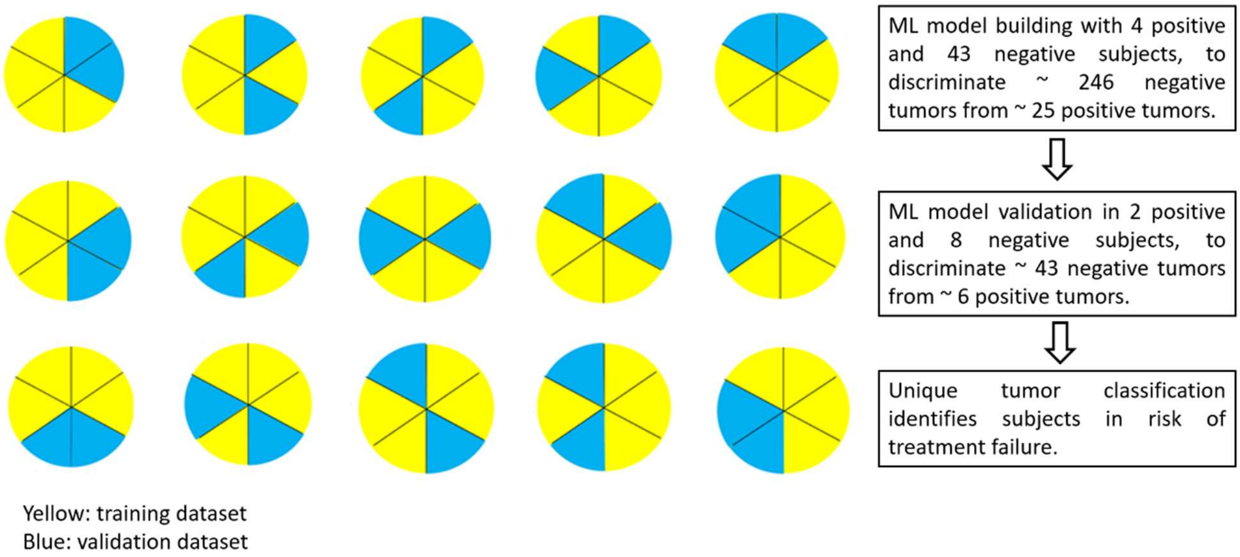
28. Orlhac F, Boughdad S, Philippe C, et al. A post-reconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018;2(Md):jnumed.117.199935. doi:10.2967/jnumed.117.199935
29. Sollini M, Cozzi L, Antunovic L, Chiti A, Kirienko M. PET Radiomics in NSCLC: State of the art and a proposal for harmonization of methodology. Sci Rep. 2017;7(1):1-15. doi:10.1038/s41598-017-00426-y

Contact:
Marcos Machado
Nuclear Medicine Department, Hospital São Rafael, Avenida Sãi Rafael, 2152, São Marcos, Salvador – BA, 41253-190, Brazil.
Tel. +55 71 3281-6891. Fax. +55 71 3281-6327.
E-mail: machado@radtec.com.br

Supplementary Material

Illustration and detailed model preparation:

The following illustration represents every ML 15-fold cross-validation. From the six groups, every subset has one subject classified as "progressive disease". The training was performed with four subsets and the validation with two subsets, where each fold randomly distributed 85% of "non-progressive disease" subjects into the training dataset and the remaining 15% into the validation datasets. Tumors from the validation datasets were scored at every fold.



ML model building with 4 positive and 43 negative subjects, to discriminate ~ 246 negative tumors from ~ 25 positive tumors.

ML model validation in 2 positive and 8 negative subjects, to discriminate ~ 43 negative tumors from ~ 6 positive tumors.

Unique tumor classification identifies subjects in risk of treatment failure.

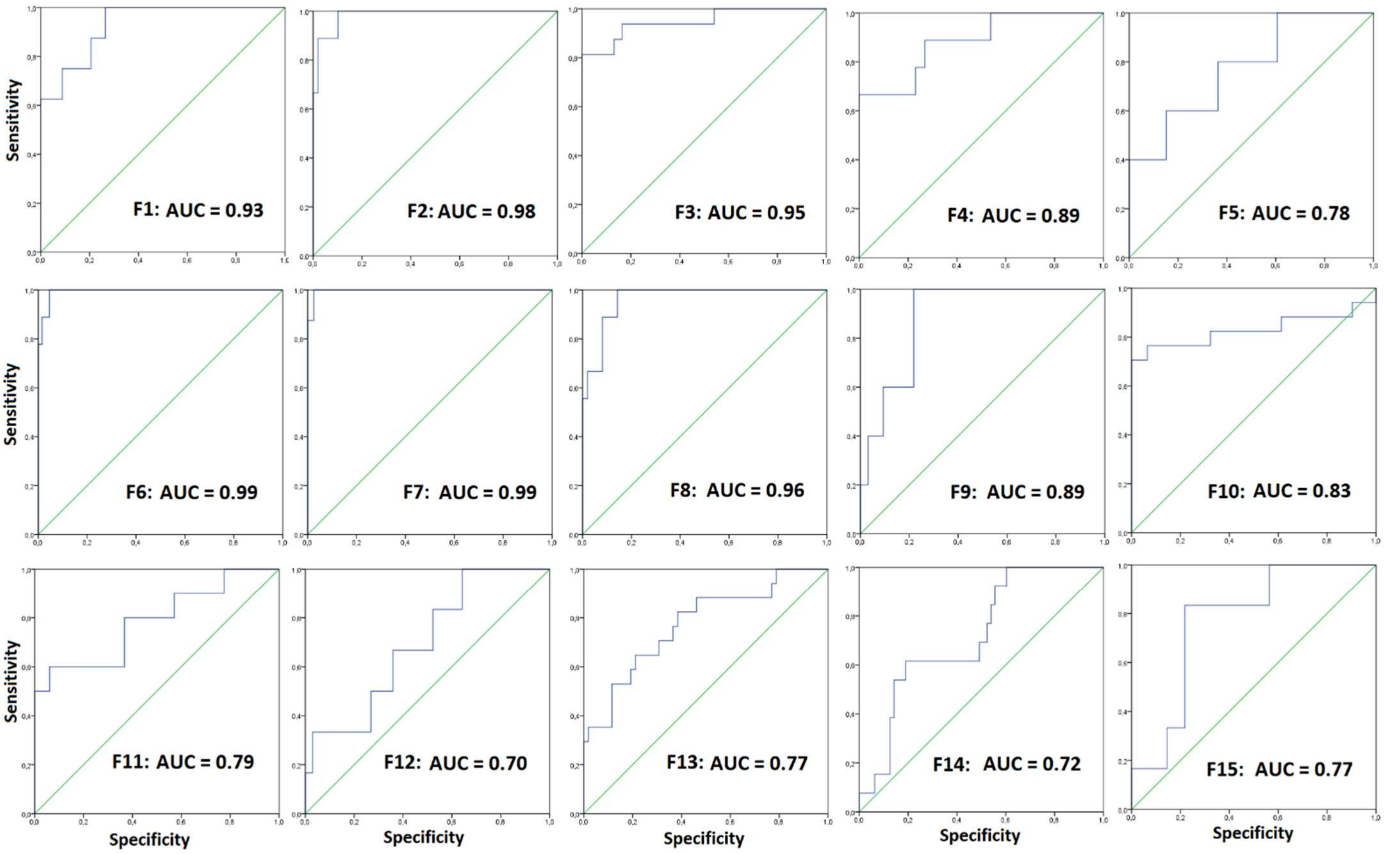
Yellow: training dataset
Blue: validation dataset

This random sampling with a replacement scheme resulted in 15 independent validation datasets at the tumor level. The model details are as follows:

RandomForestClassifier (n_estimators=1, max_depth=2, min_samples_split=20, random_state=6, class_weight='balanced_subsample')

AdaBoostClassifier (base_estimator=RandomForestClassifier, n_estimators=300, algorithm="SAMME.R", learning_rate=0.7, random_state=32)

The figure below shows the ROC curves at every validation dataset (F1 to F15).



Tumor scores identify subjects at risk of treatment failure (*S* = 0.55). The table below shows the classification performance at the subject level:

Fold	Sensitivity	Specificity
1	100%	100%
2	100%	100%
3	100%	100%
4	100%	100%
5	100%	100%
6	100%	80.0%
7	100%	100%
8	100%	100%
9	50.0%	75.0%
10	50.0%	75.0%
11	50.0%	80.0%
12	100%	60.0%
13	50.0%	100%
14	50.0%	75.0%
15	50.0%	100%
Mean (95% CI)	80.0% (43.7-97.0%)	88.3% (76.6-94.5%)