Computer Methods and Programs in Biomedicine 159 (2018) 23–30

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Radiomics-based features for pattern recognition of lung cancer histopathology and metastases

José Raniery Ferreira Junior[a,*], Marcel Koenigkam-Santos[b], Federico Enrique Garcia Cipriano[b], Alexandre Todorovic Fabro[b], Paulo Mazzoncini de Azevedo-Marques[b]

[a] São Carlos School of Engineering, University of São Paulo, São Carlos, SP 13566-590, Brazil
[b] Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP 14049-900, Brazil

ARTICLE INFO

Article history:
Received 26 July 2017
Revised 16 February 2018
Accepted 22 February 2018

Keywords:
Lung cancer
Metastasis prediction
Pattern recognition
Quantitative image analysis
Radiomics

ABSTRACT

Background and Objectives: lung cancer is the leading cause of cancer-related deaths in the world, and its poor prognosis varies markedly according to tumor staging. Computed tomography (CT) is the imaging modality of choice for lung cancer evaluation, being used for diagnosis and clinical staging. Besides tumor stage, other features, like histopathological subtype, can also add prognostic information. In this work, radiomics-based CT features were used to predict lung cancer histopathology and metastases using machine learning models. Methods: local image datasets of confirmed primary malignant pulmonary tumors were retrospectively evaluated for testing and validation. CT images acquired with same protocol were semiautomatically segmented. Tumors were characterized by clinical features and computer attributes of intensity, histogram, texture, shape, and volume. Three machine learning classifiers used up to 100 selected features to perform the analysis. Results: radiomics-based features yielded areas under the receiver operating characteristic curve of 0.89, 0.97, and 0.92 at testing and 0.75, 0.71, and 0.81 at validation for lymph nodal metastasis, distant metastasis, and histopathology pattern recognition, respectively. Conclusions: the radiomics characterization approach presented great potential to be used in a computational model to aid lung cancer histopathological subtype diagnosis as a "virtual biopsy" and metastatic prediction for therapy decision support without the necessity of a whole-body imaging scanning.

1. Introduction

Lung cancer accounts for one-third of all cancer-related deaths in the United States, with the highest mortality of all cancers [1]. The prognosis of lung cancer is still poor and varies markedly according to tumor staging at diagnosis. Tumor stage at presentation, as designated by the tumor-node-metastasis system (TNM - describes the anatomical extent of disease based on assessment of three components: extent of the primary tumor (T), absence or presence and extent of regional lymph node metastasis (N), and absence or presence of distant metastasis (M)), is the most important prognostic factor and may determine therapy [2].

However, studies showed that clinical decision making may be influenced by other tumor aspects, such as, the histopathological subtype of the lesion [3]. In clinical practice, lung cancer can be classified in two main categories: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC comprehends 85% of the cases and is mainly subclassified in adenocarcinoma (ADC), squamous cell carcinoma (SCC), and large cell carcinoma (LCC).

Besides tumor histology and staging, it has been described that other computed tomography (CT) features may also influence prognosis and response to therapy [2,4,5]. However, those features are typically described subjectively, qualitatively or semi-quantitatively, e.g. non-solid, semi-solid, or solid nodules; and single 2D measure of greatest diameter on axial plane. Furthermore, subtypes of ADC, SCC, and SCLC may present visual CT features that are similar to benign lesions [6].

Computer-aided detection/diagnosis (CAD) tools have been developed to aid specialists interpret medical imaging findings and identify early diseases, especially breast and lung tumors [7–9]. The purpose of CAD is to improve the accuracy and consistency of medical image diagnosis through computational support used as reference [10]. Traditionally, CAD systems provide a single answer (presence of a cancer, for instance) as a second opinion to special-

* Corresponding author.
E-mail address: jose.raniery@usp.br (J.R. Ferreira Junior).

https://doi.org/10.1016/j.cmpb.2018.02.015
0169-2607/

ists. They have not been developed to provide prognostic data and aid decision making of therapy.

On the other hand, radiomics is an extension of CAD that converts imaging data into a high dimensional feature space, which may ultimately correlate with clinical outcomes [11,12]. Moreover, radiomics involves extracting image features and combining them with other patient data, as available, to increase the power of decision support models [9]. In other words, radiomics is a data correlation process that recognizes diagnostic and prognostic outcome patterns based on qualitative and quantitative features combined with patient clinical descriptors.

Several qualitative and semi-quantitative (attenuation, heterogeneity, spiculation, location, size, shape, margins, calcification, cavitation and so on) and quantitative (shape, gray-level intensity, histogram, cooccurrence matrix texture, run-length matrix texture, wavelet decomposition, standardized uptake value statistics and so on) image features have been used to characterize malignant lung tumors [2,4,5,11–17]. However, a gold standard radiomic pattern for tumor characterization remains challenging.

In this context, the goal of this work is to assess the correlation of radiomics-based features obtained in CT images with tumors histology and presence of lymph node and distant metastases. Our main purpose is to build a radiomics model that potentially may aid specialists on lung cancer histological subclassification and decision making of therapy based on metastases status. In this work, we also aim to compare the performance of fine-tuned recognition of metastatic and histopathological patterns using different feature categories (clinical, computer, and both combined), machine learning classifiers (a probabilistic, an instance-based, and an artificial neural network), and (balanced and unbalanced) datasets for training.

The remainder of this paper is organized as follows: Section 2 describes material and methods used in this work, or more specifically, the radiomics processes of image acquisition in Section 2.1, image segmentation in Section 2.2, feature extraction in Section 2.3, feature selection in Section 2.4, and tumor classification in Section 2.5. Results are presented in Section 3 and discussed in Section 4. Finally Section 5 concludes this article.

2. Material and methods

2.1. Image acquisition

Our institutional research board approved this retrospective study with a waiver of patients' informed consent. We analyzed 68 (52 for testing and 16 for validation) malignant lung tumors with histology confirmed by biopsy or surgical resection (Tables 1 and 2 of supplementary material). Thin-slice contrast-enhanced CT images were acquired in a multidetectors CT scanner, after intravenous administration of iodinated contrast media, using volumetric acquisition with slice thickness and reconstruction interval of 1–1.5 mm.

Our lung cancer tumor database has 90 cases, but 18 of them were not acquired with the standard contrast-enhanced CT protocol, which would influence in image characterization process, 2 of them did not present all clinical data, and 2 of them presented other opacities attached to the tumor, which would influence in the image segmentation process. Therefore, we excluded those 22 cases from the analysis.

2.2. Image segmentation

Manual segmentation of tumors in medical imaging is a labor-intensive and time-consuming task, which limits the amount of cases that can be processed [18,19]. Hence all 68 tumors were semiautomatically segmented by the 3D region growing GrowCut algorithm from the medical image analysis and visualization Slicer platform v4.3.1 (Fig. 1) [20]. GrowCut was first validated with glioblastoma multiforme in magnetic resonance imaging [21] and then with lung tumors in CT scans [22].

In this work, we first marked two regions, one inside and one outside the tumor (Fig. 1(a)), on three slices of the CT exam, one slice for each anatomical plane, using a lung window with level of -500 and width of 1400. After that, an interactive region growing procedure based on cellular automaton detected tridimensionaly the tumor tissue (Fig. 1(b)), by labeling voxels of a convex hull of the regions with a 5% additional margin [21]. Voxel labeling is done using a weighted similarity score, which is a function of the neighboring voxel weights, and continues iteratively until a stable configuration is reached when modification of the voxel labels is no longer possible. An unlabeled voxel is labeled corresponding to the neighboring voxels that have the highest weights [22]. Next we removed the outside mark of the tumor (Fig. 1(c)), and modeled the boundary outline (Fig. 1(d)). Finally the tumor outline was exported as a Digital Imaging and Communications in Medicine radiation therapy (DICOM-RT) structure set file [23], to be used in the extraction of computer features, as follows.

2.3. Feature extraction

Tumors were characterized by 2277 quantitative features extracted from segmented images. Features of gray-level intensity, histogram, cooccurrence matrix (COM), run-length matrix (RLM), neighborhood intensity difference matrix (NIDM), Laplacian of Gaussian (LOG) filtered statistics, and shape were extracted by the IBEX radiomics platform v1.0 [24]. Histogram, COM, RLM, and NIDM features were calculated on 8-bit converted CT images to prevent sparsely populated matrices and histogram from being produced. 8-bit image quantization may also ultimately reduce the effect of potential noise in CT for soft tissue and tumor on the texture features [15–17,25].

2.3.1. Intensity features

Gray-level intensity features describe the distribution of values of individual voxels from a volume of interest (VOI) without concern for spatial relationships [9]. A total of 53 first-order intensity features were calculated in this work: energy, global maximum, global mean, global median, global minimum, global standard deviation, inter quartile range, kurtosis, local entropy maximum, local entropy mean, local entropy median, local entropy minimum, local entropy standard deviation, local range maximum, local range mean, local range median, local range minimum, local range standard deviation, local standard deviation maximum, local standard deviation mean, local standard deviation median, local standard deviation minimum, local standard deviation standard deviation, mean absolute deviation, median absolute deviation, percentile, quantile, range, root mean square, skewness, and variance [11,24]. Percentile features varied according to the interval of 5–95 with an incremental of 5, and quantile of 0.025, 0.25, 0.50, 0.75, and 0.975. Local entropy features used neighborhood of size 9 pixels, and local range and local standard deviation features used neighborhoods of size 5 pixels [17].

2.3.2. Histogram features

Gray-level histogram features are calculated from a first-order histogram that represents a particular VOI by tabulating the number of voxels within a particular value (in this work, 256 bins of 16 Hounsfield units) [16]. A total of 51 histogram features were used: entropy, uniformity, inter quartile range, kurtosis, mean absolute deviation, median absolute deviation, percentile, percentile area, quantile, range, and skewness [24]. Percentile and percentile