Computers in Biology and Medicine 91 (2017) 222–230

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Deciphering unclassified tumors of non-small-cell lung cancer through radiomics

Maliazurina Saad, Tae-Sun Choi *

School of Mechatronics, Gwangju Institute of Science and Technology, 123 Cheomdan-gwagiro, Oryong-Dong, Buk-gu, Gwangju 61005, South Korea

ARTICLE INFO

Keywords:
Non-small-cell lung cancer
Subtyping
Well differentiated
Poorly differentiated
Radiomics

ABSTRACT

Background: Tumors are highly heterogeneous at the phenotypic, physiologic, and genomic levels. They are categorized in terms of a differentiated appearance under a microscope. Non-small-cell lung cancer tumors are categorized into three main subgroups: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. In approximately 20% of pathology reports, they are returned unclassified or classified as not-otherwise-specified (NOS) owing to scant materials or poor tumor differentiation.

Method: We present a radiomic interrogation of molecular spatial variations to decode unclassified NOS tumor architecture quantitatively. Twelve spatial descriptors with various displacements and directions were extracted and profiled with respect to the subgroups. The profiled descriptors were used to decipher the NOS tumor morphologic clues from the imaging phenotype perspective. This profiler was built as an extended version of a conventional support-vector-machine classifier, wherein a genetic algorithm and correlation analysis were embedded to define the molecular signatures of poorly differentiated tumors using well-differentiated-tumor information.

Results: Sixteen multi-class classifier models with an 81% average accuracy and descriptor subset size ranging from 12 to 144 were reported. The average area under the curve was 86.3% at a 95% confidence interval and a 0.03–0.08 standard error. Correlation analysis returned an unclassified NOS membership matrix with respect to the cell-architecture similarity score for the subgroups. The best model demonstrated 53% NOS reduction.

Conclusion: The membership matrix is expected to assist pathologists and oncologists in cases of unresectable tumors or scant biopsy materials for histological subtyping and cancer therapy.

1. Introduction

Most cells of the body are differentiated, which provides them with unique signatures based on their specialization, such as those of the kidney-filtering mechanism and heart muscles. However, cancer cells mutate. As they grow and divide, these cells tend to make erroneous copies of their DNA, thereby leading to further mutations. Such mutations comprise the underlying cause of heterogeneity. Cancers are therefore comprised of cells that may appear like the cells from which they originated (well differentiated), or they may appear nothing like the cells from which they arose (poorly differentiated) [1].

Non-small-cell lung cancer (NSCLC) is a type of lung cancer grouped based on the clinical behavior and histological appearance of the cancer cells. There are subtypes of NSCLC that basically define the different types of lung cells from which they originated. Adenocarcinoma (ADC) is usually found in the outer parts of the lung, either peripherally located in the smaller airways or centrally located in the main bronchus. Squamous cell carcinoma (SCC) generally arises within the lungs, inside a large bronchus. Large-cell carcinoma (LCC) is an undifferentiated tumor that lacks diagnostic features, and therefore a default classification, to some extent, which is made when other specific histologies have been excluded [2].

Aside from these three common classification subgroups, 20% of pathologist reports are returned as not-otherwise-specified (NOS) tumors. This classification status arises mainly because of inadequate tissue; e.g., the biopsy material was scant or the tumor was poorly differentiated. NSCLC tumors are known to be very heterogeneous. This means that in an area that appears predominantly SCC, a centimeter away, for example, may appear predominantly ADC. Hence, a pathological examination based on small biopsy or cytological samples may not be effective in defining the subgroups because it does not represent the cell activity or toxicity profile of the whole tumor. In such a scenario,

* Corresponding author.
E-mail addresses: eena@gist.ac.kr (M. Saad), tschoi@gist.ac.kr (T.-S. Choi).

https://doi.org/10.1016/j.compbiomed.2017.10.029
Received 17 August 2017; Received in revised form 24 October 2017; Accepted 24 October 2017

misdiagnoses may occur. This is the underlying reason why a pathologist may choose to define a tumor as NOS. Misclassification of subgroups may cause inappropriate therapeutic options to be administered, thus leading to complications. For example, it has been reported that certain drugs, such as ALIMTA (pemetrexed), can induce fatal bleeding in SCC patients and should thus be avoided.

As the field of oncology has evolved, the methods of diagnosing a tumor subgroup have also advanced. Pathologists previously identified cell activities using the naked eye under a normal microscope or using a light microscopy technique. The lack of a differentiated tumor, on the other hand, requires a more sophisticated method, such as immunohistochemistry (IHC), which employs monoclonal antibodies directed against very specific epitopes on the surface and sometimes in the nuclei of the target cells, such as TTF-1 for ADC and P63 for SCC [3–5].

However, even the most sophisticated method, such as IHC, is still impractical in the event of unresectable tumors or scantly available tissue, which may hamper the process of precisely subtyping a tumor. This has prompted the concept of using the quantitative analysis of a radiomic approach to quantify tumor heterogeneity through image processing and data-mining algorithms of routinely acquired computed tomography (CT) images to decode the unclassified NOS tumors in terms of their cell architectures.

The '-omics' studies have been long established in research fields, such as proteomics, genomics, and metabolomics. These relate to the study of protein, gene, and metabolic structures, respectively. Radiomics, on the other hand, is an emerging field that converts radiography data into a high-dimension, mineable feature space using data characterization algorithms for hypothesis generation or testing [6–9]. Radiomics is a set of patterns or image structures, such as points, edges, curves, shapes, textures, and others, that are numerically represented. Tissue characteristics, such as cell density and metabolic activity, can be visualized through such diagnostic imaging.

The heterogeneity observed in a tumor is a reflection of the phenotypic variation reported to be associated with the underlying gene expression [10–14]. Radiomics has shown immense potential in aiding cancer diagnosis and prognosis in the near future. Recent studies on tumor heterogeneity as captured by tomographic imaging have revealed it as an emerging approach in disease treatment and prevention [15–17]. The objective of the present study is therefore to decipher the molecular clues or cell architectures of NSCLC-NOS tumors using the pathologically defined NSCLC subgroups of ADC, SCC, and LCC information through a quantitative analysis of radiomics. Through the present work, we strive to assist the pathologist and oncologist community in improving the specificity of NSCLC histologic classification by reducing the proportion of NOS cases.

2. Materials and methods

2.1. Database

The Lung1 dataset, publicly available in the Cancer Imaging Archive (TCIA) funded by the National Cancer Institute (NCI) was included in the present study [18]. It consists of the records of 422 NSCLC patients that were treated at the Maastricht University Medical Center, Maastricht, The Netherlands, in accordance with Dutch law and approved by the Institutional Review Boards [19]. For these patients, de-identified pre-treatment CT scans, manual delineations, clinical data, and survival data were available.

The CT examinations were acquired using a CT scanner from the manufacturers CMS and Siemens. The acquisition protocol varied slightly for different patients depending on the patient's size. Exposure settings ranged from 120 to 140 kVp with a tube current of 40–80 mAs. The reconstructed pixel resolution was 0.977 × 0.977, and the images were reconstructed at 512 × 512 pixel matrices. The slice thickness was 3.0 mm in all examinations, and the convolution kernels were B30s, B30f, B18f, B19f, and a few others. The number of cases diagnosed for the well-defined subgroups of ADC, SCC, and LCC were 51, 152, and 114, respectively.

The remaining were left unclassified (63 of NOS) or unknown (42 of NA). The inclusion criteria were patients with a measurable tumor size (>30 mm) and stage III tumors according to Tumor/Node/Metastasis (TNM) staging. The patients with metastasis were excluded. Hence, 20 patients were curated from each of the ADC, SCC, and LCC cases for the profiling process, and 28 patients were selected from NOS cases for the correlation analysis. Demographically, male patients dominated the constructed dataset, constituting overall 75% of the 88 patients. The mean ages for each subgroup were 61, 67, for ADC and SCC, respectively, and 64 for both LCC and NOS.

2.2. Radiomic framework

Fig. 1 illustrates the overall process of the radiomic modeling. Initially, a tumor is extracted using a semi-automated system in the first layer. On the middle layer, the descriptors are extracted before submission to the last layer, where they are profiled or selected based on the classification performance of the well-defined tumors. Once finalized, these descriptors facilitate the association of well-defined tumors with the unclassified tumors.

2.3. Computer-assisted tumor segmentation

A semi-automated computer-aided tumor segmentation method was employed, as shown in the upper inset of Fig. 1. The contrast-enhanced CT images first underwent automated thoracic segmentation using adaptive thresholding [20] and connected component labeling [21,22]. Lung region extraction and nodule candidate detection were performed using automated seeded region growing [23]. A radiologist with more than five years of experience supervised the tumor extraction process by marking and finalizing the target region.

2.4. Descriptor extraction

A second-order statistical texture descriptor, as studied by Elmoez et al. [24] and called a three-dimensional (3D) co-occurrence matrix, was used mainly because it considers the spatial dependence among slices in a z direction (volumetric data). Similar to a traditional two-dimensional (2D) co-occurrence matrix, it accumulates the number of pixel pairs having intensities i and j, and $P(i, j)$. However, this matrix is defined along a displacement $d = (dx, dy, dz)$, where dx and dy are similar to the 2D case, while dz represents the number of pixels moving along the z-axis. Twelve statistical features $f = \{f_1, f_2 \ldots. f_{12}\}$ quantifying the spatial dependency were extracted from this matrix, as demonstrated in the Supplementary Table S1.

Spatial dependency extraction requires displacement d and direction θ of a particular measurand (e.g., a pair of pixels) to be considered. The more options of d and θ that are considered, the higher is the computational complexity. However, since it can never be known which pair of pixels for which displacement and direction will give the most predominant information, all the possibilities must be considered. We thus employed four displacements and 13 directions, as shown in the Elmoez work, resulting in the feature vector of size 624.

2.5. Descriptor profiling and membership score assignment

As illustrated by the lowest inset of Fig. 1, the last section is divided into sub-phases: descriptor selection (genetic algorithm (GA)), multi-class classification (support vector machines (SVMs)), and membership score assignment (correlation analysis). It was built as an extended version of a classic classification framework with the descriptor selection and correlation analysis embedded as a pre- and post-framework. The outcome of the overall architecture was a membership matrix (M) that quantifies the similarity score between an unclassified tumor and the