Journal of Biomedical Informatics: X 5 (2020) 100067

Contents lists available at ScienceDirect

Journal of Biomedical Informatics: X

journal homepage: www.journals.elsevier.com/journal-of-biomedical-informatics-x

A spatial neighborhood methodology for computing and analyzing lymph node carcinoma similarity in precision medicine

T. Luciani[a], A. Wentzel[a], B. Elgohari[b], H. Elhalawani[b], A. Mohamed[b], G. Canahuate[c], D.M. Vock[d], C.D. Fuller[b], G.E. Marai[a,*]

[a] Department of Computer Science, University of Illinois at Chicago, United States
[b] MD Anderson Cancer Center, United States
[c] Department of Computer Science, University of Iowa, United States
[d] Department of Biostatistics, University of Minnesota, United States

ARTICLE INFO

Keywords:
Precision medicine
Spatial similarity
Topology
Visual interface
Oncology

ABSTRACT

Precision medicine seeks to tailor therapy to the individual patient, based on statistical correlates from patients who are similar to the one under consideration. These correlates can and should go beyond genetics, and in general, beyond tabular or array data that can be easily represented computationally and compared. For example, in many types of cancer, cancer treatment and toxicity depend in large measure on the spatial disease spread—e.g., metastasizes to regional lymph nodes in head and neck cancer. However, there is currently a lack of methodology for integrating spatial information when considering patient similarity. We present a novel modeling methodology for the comparison of cancer patients within a cohort, based on the spatial spread of the lymph nodes affected in each patient. The method uses a topological map, bigrams, and hierarchical clustering to group patients based on their similarity. We compare this approach against a nonspatial (categorical) similarity approach where patients are binned solely by their affected nodes. We present similarity results on a 582 head and neck cancer patient cohort, along with two visual abstractions for analysis of the results, and we present clinician feedback. Our novel methodology partitions a patient cohort into clinically meaningful groups more susceptible to treatment side-effects. Such spatially-aware similarity approaches can help maximize the effectiveness of each patient's treatment.

1. Introduction

The United States National Cancer Institute estimates that more than 51,000 people in the United States were diagnosed in 2018 with head and neck squamous cell carcinoma (HNSCC) [1]. Of these HNSCC cases, more than 90% result as oropharyngeal carcinomas (OPC), which include cancers of the larynx (voice box), pharynx (throat), lips, tongue, and nose [2,3]. At the same time, the large number of HNSCC cases makes possible the creation of big data repositories consisting of the demographic and clinical characteristics, treatments, and outcomes of patients undergoing therapy [4]. These repositories present opportunities towards informing and further personalizing treatment on a per-patient level, rather than relying on clinician experience or institutional memory alone [5–7]. Under a healthcare model termed ”precision medicine”, clinicians aim to use these patient repositories to tailor therapy decision to the individual patient, based on data from patients who are similar to the one under consideration [8,9]. Currently, these correlates typically include age, performance status, clinical staging information, and sometimes genetics—attributes that can be statistically aggregated, matched and analyzed.

Yet, similar to most other cancer types, HNSCC treatment and side effects depend in large measure on the spatial location and spread of the cancer. In particular, for more than 50% of OPC patients, the treatment and side-effects are heavily influenced by the spread of disease to lymph nodes (LN) and their corresponding areas (levels), at risk for metastases. OPC generally metastasizes to regional LNs following the lymphatic drainage of the head and neck [10], often resulting in chains of affected LNs along the drainage pathway. These chains correspond to the spread of disease to specific locations of the head and neck and are thus defined by their spatial attributes. Therefore, for those patients receiving intensity-modulated radiation therapy (IMRT), these chains represent additional targets that must receive radiation treatment. Further complicating matters, the soft tissue structures of the head and neck (organs, muscles, etc.) are highly susceptible to both direct and

[*] Corresponding author.
E-mail address: gmarai@uic.edu (G.E. Marai).

https://doi.org/10.1016/j.yjbinx.2020.100067

Available online 03 February 2020

indirect radiation exposure [1], and the increased toxicity to specific regions has been shown to correlate with post-therapy quality of life [8]. For example, aspiration and dysphagia side-effects affect as many as 30%-50% of patients treated with IMRT [11]. Therefore, clinicians believe that grouping patients by their patterns of nodal spread can help improve treatment strategies regarding both efficacy and toxicity.

The state of the art in lymph pattern similarity uses either categorical (i.e., nonspatial) matching of node labels, or relies on clinician memory. The first approach does not capture the spatial patterns of disease spread, and the second approach clearly does not scale well. Because within a patient cohort there are many rare or unique combinations of spatial affected chains, analyzing and interpreting the results of any lymph similarity measure is further challenging. Precision medicine stands to benefit from scalable, rigorous computing methodology that takes into account both the information about metastasized nodes and about the pathways that connect them, and facilitates the analysis and interpretation of the resulting similarity measures.

At the same time, spatial similarity has been facilitated in many domains such as mechanical engineering [12], bioinformatics [13], and oncology [14,15] by encoding spatial relationships through either topology-based or shape-based techniques. These techniques have the ability to ”exhibit common classes of descriptive spatial (topological) features that are quantified by definition of computable measures” [16]. Both topology and shape-based techniques aim to extract spatial attributes, then establish a relationship between corresponding attributes in different patients. However, shape-similarity based methods tend to focus on classifying models of very different shapes, and fall short of distinguishing anatomical objects within the same class unless the objects have easily identifiable structures, such as the mandible and outer body contour [14,17,18]. In the case of lymph nodes, structures are in the same class and do not have easily identifiable features. However, OPC patient analysis presents an opportunity for topology-based techniques.

In this methods paper, we present a novel topology-based modeling methodology for the comparison of patients within a cohort, based on the spatial pattern of lymph nodes affected by disease. As part of this methodology, we construct a topological map, we define computational representations, and we introduce a novel graph-based measure to derive patient LN spread similarity. We further construct a novel visual interface to interpret the spatial similarity results, followed by a novel dendrogram visual encoding to communicate the results to clinicians. We use a cohort of 582 post-therapy OPC patients to evaluate the benefits of this methodology over the nonspatial approach, and to illustrate its potential for clinical application. We first contrast the novel spatial measure results against the results obtained using a nonspatial (categorical) labeling of the nodes. We then hypothesize that the underlying spatial information contained within the chains of affected LN levels would significantly correlate with post-therapy side-effects known to arise due to radiation toxicity. This novel computing methodology should further allow for binning of patients in cohorts deemed by clinicians as significantly more informative than nonspatial (categorical) binning.

2. Materials and methods

2.1. Method overview

Our methodology is constructed as follows (Fig. 1): the LN levels for eligible patients are manually segmented from contrast-enhanced computed tomography imaging data. We then construct a LN topological map, based on the level location and its surrounding local neighborhood, and using the medical literature [19] and clinician input [20]; because of left-right symmetry in the human head and neck, this is a 2D map with cells for each node region. To facilitate patient comparison using the spatial information, we next define and construct a dual-graph representation over the topological map; this representation captures the neighbor relationships among the lymph nodes. We use a novel graph-based representation and spatial measure to compute the pairwise similarity between patients. Next, we perform hierarchical agglomerative clustering and visual analysis on the similarity output and compare the resulting patient groupings. The results are then presented to the clinicians for interpretation of the rankings and clusters of patients. Finally, we perform a statistical analysis to determine if our spatial measure is significantly correlated with post-treatment toxicity outcomes. We describe below in detail each component of this method.

2.2. Patient cohort

Oropharyngeal cancer (OPC) patients who were treated at the MD Anderson Cancer Center between 2005 and 2013 were retrospectively reviewed under an approved IRB protocol. Out of the 644 eligible patients who had a pathologically proven OPC, either with a positive biopsy or a surgical excision and received treatment (i.e., radiotherapy +/- chemotherapy) with a curative intent, 582 patients had affected lymph nodes and were included in this study. Affected lymph node (LN) levels were collected from contrast enhanced computed tomography (CECT) diagnostic scans which took place at patients’ initial visit for staging and disease assessment. LN levels (retropharyngeal (RP), submental (Ia), submandibular (Ib), upper, medial and lower jugular (II, III, IV respectively) and level V a, b) were defined based on anatomical landmarks and were coded in relation to tumor position. Patients’ relevant demographic, clinical, and toxicity data (toxicity of interest were feeding tube and aspiration at six months) were retrieved from electronic medical records. Only the LN information was used as input to our method.

Table 1 shows the post-therapy side-effect counts and patient characteristics across the cohort. Of the 582 patients who underwent intensity-modulated radiotherapy, 163 patients suffered from either post-therapy dysphagia side-effects, with 95 (16.32%) patients reporting aspiration (breathing a foreign material to the airways, such as saliva) and 99 (17.01%) requiring a feeding tube six months after the end of radiotherapy treatment (Feeding Tube at 6 months).

2.3. Topological map

To enable spatial comparison, we first defined and constructed a novel 2D topological map over the LN levels, based on the consensus guidelines for the delineation of the head and neck [19], and using the left-right symmetry of the human head and neck, as well as input from our clinician collaborators. Each cell in this topology, shown in gray in Fig. 2 (right), corresponds to an LN level in the human head and neck, based on the spatial location and local neighborhood of each level. Over this topology, we then defined a dual graph representation, shown in red in Fig. 2 (right), where each cell was represented as a node in an undirected graph, and edges were created between each pair of adjacent faces. Using this novel abstraction, a chain of disease spread would follow the links between the adjacent faces; for example, the path connecting LN levels 2B-2A-3 corresponds to a lymph chain of spread. We decided to place the Retropharyngeal (RP) LN, a LN group near the base of the skull, as a disconnected node in the graph (upper left) because metastasis to this group bears a poor prognosis to OPC patients and requires specialized treatment.

Next, the graph template (Fig. 2 Left) was encoded as an adjacency matrix (Fig. 3). In the adjacency matrix M, each row and column correspond to one of the LN levels in the graph, whereas individual cells encode the edge information (a.k.a., adjacency) in the graph, as follows:

$$M(i,j) = \begin{cases} 1, & \text{if } LN_i \text{ and } LN_j \text{ are connected by an edge OR } i = = j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

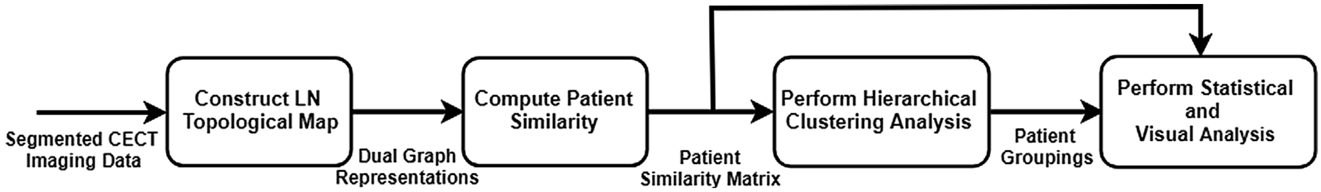


Fig. 1. Pipeline detailing the steps and data flow of our presented methodology. After receiving the contrast-enhanced computed tomography (CECT) images from the clinicians, we construct a topological mapping of each patient's affected nodes and the connections between them. The result matrices are used to compute similarity using a distance coefficient; hierarchical clustering is performed on the ranked patient scores to determine patient groups; statistical and visual analysis is performed on the groups to determine groups with higher toxicity outcome rates, and validate the results.

Table 1
Patient Characteristics and Post-therapy Side Effects.

Characteristics		N (%)
Post-therapy Side Effect		
	Feeding tube at 6 mo.	99 (17.01%)
	Aspiration	95 (16.32%)
	No side effect	388 (66.67%)
Gender		
	Male	512 (87.97%)
	Female	70 (12.03%)
T-category (T)		
	Tx	1 (0.17%)
	Tis	1 (0.17%)
	T1	129 (22.16%)
	T2	245 (42.10%)
	T3	121 (20.79%)
	T4	85 (14.61%)
N-category (N)		
	N1	72 (12.37%)
	N2	492 (84.54%)
	N3	18 (3.09%)

where LN_i and LN_j are lymph node levels in the graph.

Since the RP LN level appears as a disconnected node on the graph, we handle it as a special case. Therefore, the resulting matrix M has dimensions of 9x9, for the nine groups of lymph nodes that are connected in the graph representation. The adjacency matrix is by construction symmetric.

$$\begin{array}{c|ccccccccc|}
 & 1A & 1B & 2A & 2B & 3 & 4 & 5A & 5B & 6 \\
\hline
1A & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1B & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
2A & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
2B & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
3 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\
4 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\
5A & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
5B & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
6 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1
\end{array}$$

Fig. 3. Adjacency matrix encoding edge connections between nodes in the topological map. Rows and columns in this matrix correspond to the LN nodes in the map. A matrix element is either 1 when there is an edge connecting its row LN with its column LN in the map, or 0 when there is no connection. For clarity, main diagonal elements are set to 1 here (all nodes are self-connected), but could alternatively be used to encode the involvement status of each LN.

2.4. *Patient spatial data encoding*

Using the adjacency matrix thus defined, we then encode the LN spatial spread of disease, or lymph node *involvement*, for each patient. From a radiation oncology perspective, neck lymph nodes are classified and treated on the basis of the level where they are located; the location is determined through physical examination and/or radiological imaging. If at least one lymph node in a given level is affected with cancer

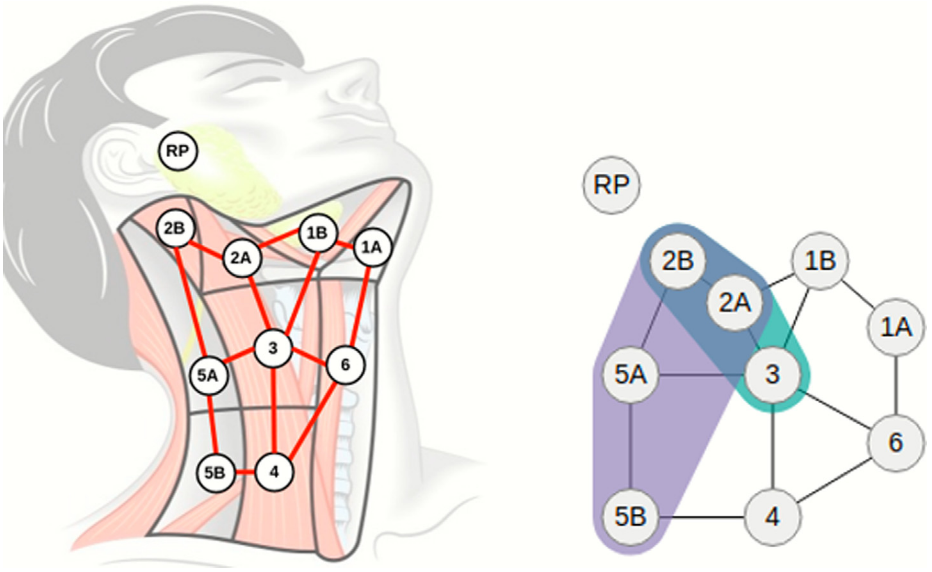
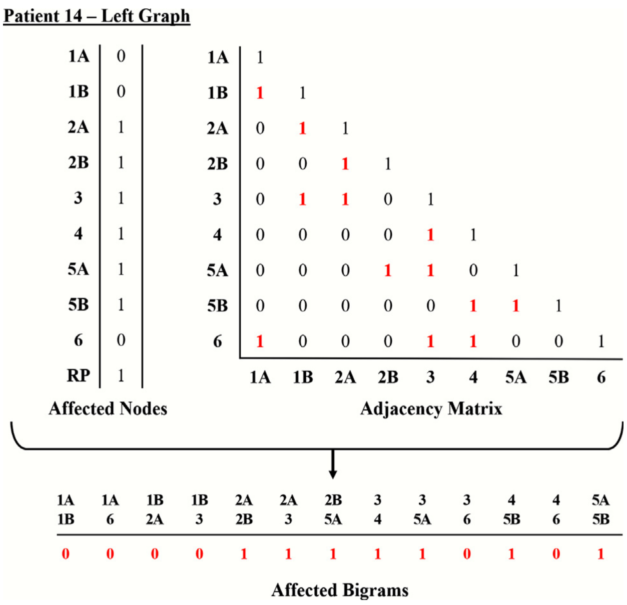


Fig. 2. Topological map and graph representation. (Left) A novel topological map was constructed over the lymph node regions (shown in gray), overlaid with a dual graph representation (red) of the map showing the connectivity between the lymph node levels. The Retropharyngeal (RP) lymph nodes are a group of nodes near the base of the skull and are disconnected from the dual graph because when affected by disease, they require specialized treatment. (Right) A compact visual representation was derived from the red graph representation to visually illustrate metastasis over both sides of the head and neck, using symmetry and color to distinguish between left (green), right (purple), and bilateral (blue) spread. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



1A	0
1B	0
2A	1
2B	1
3	1
4	1
5A	1
5B	1
6	0
RP	1

1A	1								
1B	1	1							
2A	0	1	1						
2B	0	0	1	1					
3	0	1	1	0	1				
4	0	0	0	0	1	1			
5A	0	0	0	1	1	0	1		
5B	0	0	0	0	0	1	1	1	
6	1	0	0	0	1	1	0	0	1
	1A	1B	2A	2B	3	4	5A	5B	6

1A 1B	1A 6	1B 2A	1B 3	2A 2B	2A 3	2B 5A	3 4	3 5A	3 6	4 5B	4 6	5A 5B
0	0	0	0	1	1	1	1	1	0	1	0	1

Fig. 4. Construction of chain (bigram) involvement for Patient #14 (Fig. 7 top left). First, affected nodes are identified for the specific patient (top left). Connections between nodes in the topological map, represented here by the global adjacency matrix (top right), are used to identify the 13 possible node connection. Finally, a 13-dimensional bigram vector is constructed, where each value corresponds to a 1 when both nodes connected by a given edge are affected, and 0 otherwise.

cells, radiation oncologists refer to the corresponding node level as being *involved* with disease, and they *involve* the whole node level in treatment.

Let us consider first, for clarity, the case of a patient who has nodes affected on one half of the head only; we will handle the full case of a patient with nodes affected on both sides of the head via symmetry. To encode the spatial involvement of lymph nodes for such a patient, we construct a vector based on the involvement status of that patient's LN levels, and using the adjacency matrix, as illustrated in Fig. 4. First, we consider the patient's involved LN levels and construct a vector of affected node levels, as follows:

$$\mathrm{N}_p(i) = \begin{cases} 1, & \text{if } LN_i \text{ is involved in patient } p \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathrm{N}_p(i)$ is the vector element that corresponds to the involvement status of the LN level LN_i for the specific patient p.

Next, to incorporate the topology information, we encode the edges to and from the involved LNs in that patient, using the template adjacency matrix. As illustrated in Fig. 4, we enumerate as a bigram label [21] each pair of LN levels connected by an edge in the template map. We then construct a bigram vector with these bigram labels. In the bigram vector, $\mathrm{B}_p(i, j)$ is the bigram vector element that corresponds to the involvement status of the (i, j) edge for that patient p, i.e., either 1 if both end node regions are involved, or 0 otherwise:

$$\mathrm{B}_p(i, j) = \begin{cases} 1, & \text{if } LN_i \text{ AND } LN_j \text{ are involved in patient } p \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

for all (i, j) pairs, $i \neq j$, for which $M(i, j) = =1$ in the adjacency matrix, and LN_i and LN_j correspond to the LN levels i and j.

We choose not to enumerate further than the two-node combinations because of the small number of nodes in the graph—if all n-grams were enumerated, the similarity distance between patients would increase, and the similarity score for partial pattern matches would decrease. Furthermore, permutations of each bigram are considered once, e.g., bigram permutations between LN levels 2A and 2B, 2A-2B and 2B-2A, are considered as being the same.

Generalizing to the full case of a patient with lymph node involvement on both sides of the head (Fig. 5), we construct one node-bigram vector for the left side of the head of patient p, and one node-bigram vector for the right side of the head. Using matrix notation, where $[x, y]$ denotes the concatenation of vectors x and y, and the N_p and B_p formulas above, the left-side node-bigram vector and the right-side node-bigram vector are:

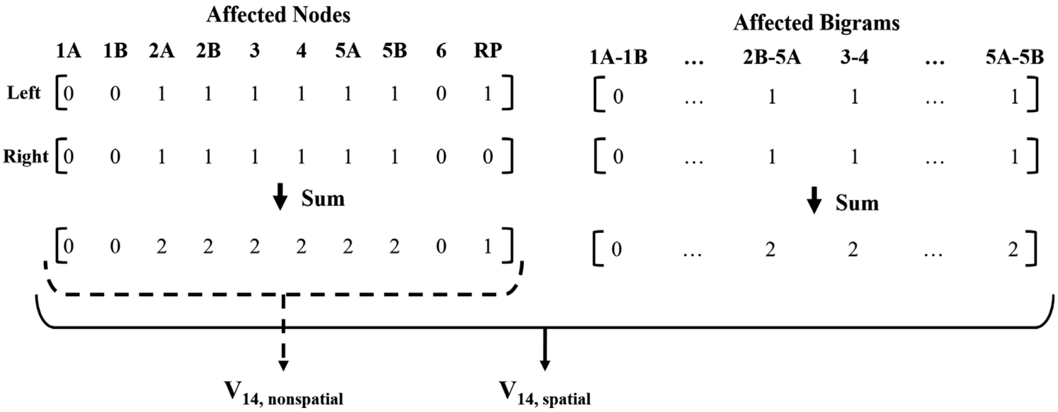
$$\mathrm{v}_{p-\mathrm{left}} = [\mathrm{N}_p, \mathrm{B}_p]_{\mathrm{left}} \tag{4}$$

$$\mathrm{v}_{p-\mathrm{right}} = [\mathrm{N}_p, \mathrm{B}_p]_{\mathrm{right}} \tag{5}$$

Because in this application the oncologists were interested in detecting both straight and symmetric similarity, i.e., also situations where one patient's left spread matches (potentially partially) another patient's right spread, we then add the left and right information into the same vector representation. Through the summation procedure, the two vectors are added, as illustrated in Fig. 5, to account for symmetry. Thus, in the general case, each patient is described by:

$$\mathrm{v}_p = \mathrm{v}_{p-\mathrm{left}} + \mathrm{v}_{p-\mathrm{right}} \tag{6}$$

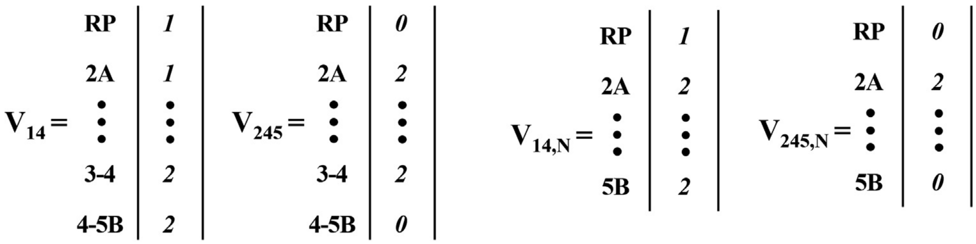
where $v_{p-\mathrm{left}}$ is the node-bigram vector for the left side of patient p, and $v_{p-\mathrm{right}}$ is the node-bigram vector that corresponds to the right side of patient p, calculated as described above. Through this concatenation and summation procedure, in total, 9 node weights and 13 bigrams weights are included into the patient vector, representing the 18 nodes and 26 bigrams on both sides of head and neck.



	1A	1B	2A	2B	3	4	5A	5B	6	RP
Left	0	0	1	1	1	1	1	1	0	1
Right	0	0	1	1	1	1	1	1	0	0
	0	0	2	2	2	2	2	2	0	1

1A-1B	…	2B-5A	3-4	…	5A-5B
0	…	1	1	…	1
0	…	1	1	…	1
0	…	2	2	…	2

Fig. 5. Construction of the spatial and nonspatial vectors for patient #14 (Fig. 7 top left). First, affected lymph nodes are encoded in a separate vector for each half of the head. Bigram involvement is then encoded into a vector for each side of the head using the matrix in Fig. 3. The final vectors are then constructed by adding the involvements for both sides of the head together to account for symmetry. Spatial vectors are constructed by augmenting the nonspatial node vector with the bigram involvement.



(a) Spatial Similarity representation

(a) Nonspatial Similarity representation

Fig. 6. An illustration of the involvement vectors v constructed for Patient #14 and Patient #245. (a) The vectors v constructed for the spatial similarity measure. (b) The vectors v constructed for the nonspatial measure. Note that the spatial vectors (a) include bigrams, and thus spatial structure, while the nonspatial vectors (b) do not.

Last, we encode the special-case RP status via two boolean flags related to the left and right involvement (Fig. 6). For later analysis, we furthermore encode the laterality of nodal involvement for each patient using the position of their primary tumor: for patients with right-sided primary tumors, right-sided LNs are encoded as 'ipsilateral' structures with tumor on the right; for patients with left-sided primary tumors, left-sided LNs are encoded as 'contralateral' structures with tumor on the left. We also encode with an additional variable the total number of bilaterally affected nodes in the patient's head. That is, a patient with nodes 2A and 2B affected on both sides of the head would have an additional variable with the value of 4. In this way, we bias the similarity measure to capture users with similar overall nodal spread that may have few true matches due to having uncommon or extensive nodal involvement.

2.5. Similarity computation

Using the vector representations derived above, we compute the LN similarity between any two patients using the Tanimoto coefficient [22]. The Tanimoto coefficient is an extension of the Cosine similarity and the Jaccard coefficient [23] for non-binary attributes. The Tanimoto coefficient is widely used in chemi-informatics [24], and also in image analysis [25], intrusion detection [26], and data mining in general [27,28]. We chose the Tanimoto coefficient based on its ability to handle non-binary data, because our folded vectors include non-binary elements (i.e., values of 2, in addition to 0s and 1s). The cohort was ranked in pairwise-fashion by computing the Tanimoto coefficient between each of the newly constructed vectors:

$$T(\mathrm{v}_p, \mathrm{v}_q) = \frac{\mathrm{v}_p \cdot \mathrm{v}_q}{\|\mathrm{v}_p\|^2 + \|\mathrm{v}_q\|^2 - \mathrm{v}_p \cdot \mathrm{v}_q} \tag{7}$$

where the function $T(\mathrm{v}_p, \mathrm{v}_q)$ returns the Tanimoto coefficient between the vectors v of patients p and q.

We note that while other distance coefficients could be used, the rankings resulting from the application of these coefficients would be similar to each other [24], given the limited range of values in our feature vectors (0,1, or 2). Thus, the use of a different coefficient would not affect the clustering results; we used Tanimoto because it was a more correct choice for our problem, and widely used in the literature.

In order to investigate whether incorporating spatial information about the lymph node chains (i.e., the spatial location and neighborhood of the nodes involved) partitioned patients more meaningfully than only considering the level itself (i.e., nonspatial labels), we next constructed a vector using only the involvement status of the LN level labels in the nonspatial (categorical) representation, and again ranked the cohort in pairwise-fashion by using the same formula. This second calculation considers each patient's LN level involvement status only (i.e., only the affected nodes in the graph representation), as opposed to a combination of status and pathways (affected nodes and edges in the graph representation).

To illustrate, in contrast, how these two approaches, spatial and nonspatial, work, let us consider patients #14 and #245 from Fig. 7 (top left). Patient #14 possesses a bilateral involvement of LN levels 2A, 2B, 3, 4, and 5B, and a unilateral involvement on one RP LN level, while Patient #245 possess a bilateral involvement of LN levels 2A, 2B, 3, and 4. Fig. 6 illustrates the corresponding vectors that are constructed for the spatial (Fig. 6a) and nonspatial (Fig. 6b) approaches. Computing the distance coefficient between both sets of patient vectors results in a similarity score of 0.87 for the spatial approach, and 0.76 for the nonspatial approach.

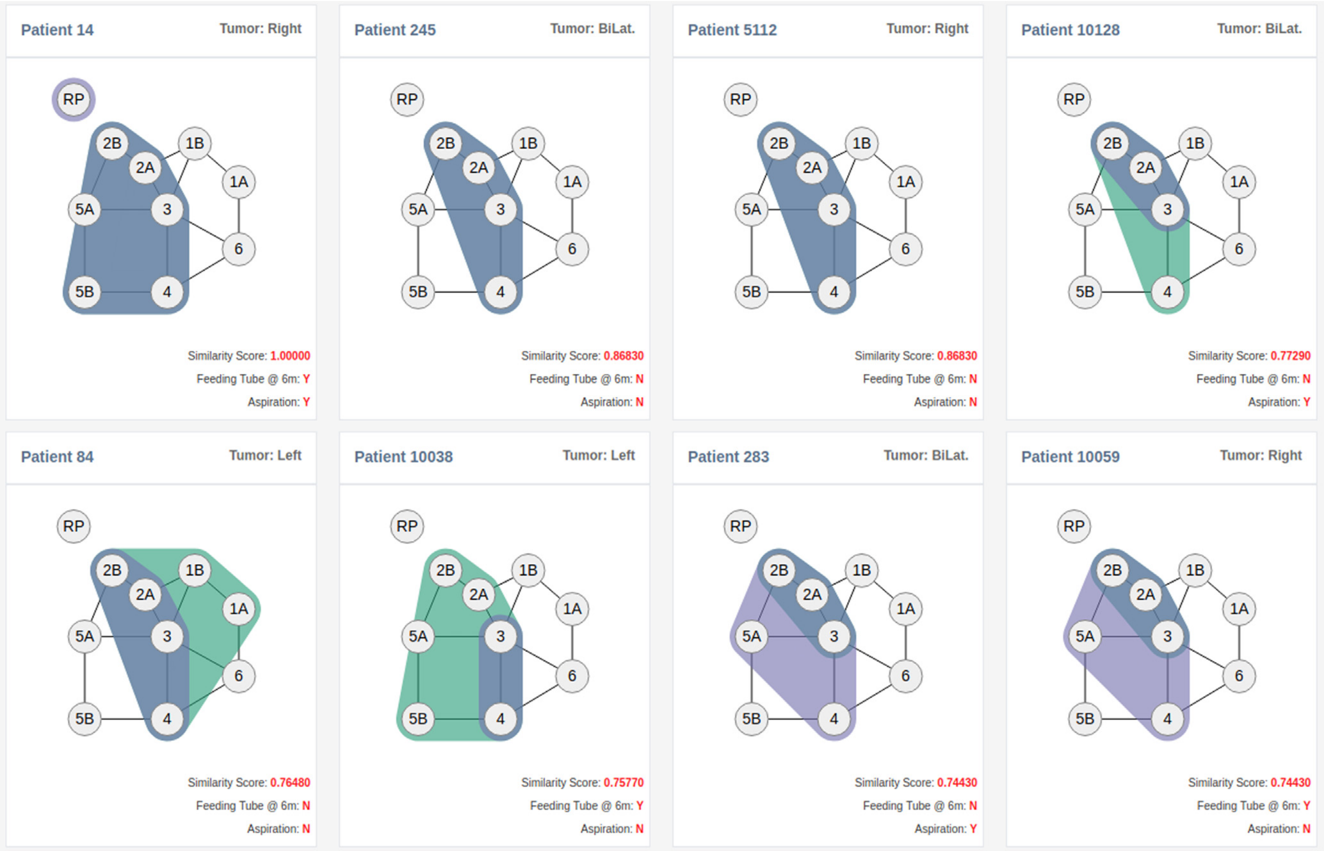
After ranking each patient, we construct two similarity matrices, one for each of the spatial and nonspatial approaches, using the similarity scores between each patient pair in the cohort. The result of this step is a similarity matrix for each approach, with the number of rows/columns in each matrix equal to the number of patients in the repository. These matrices are then used in the hierarchical clustering analysis. The patient similarity was implemented using Python 2.7.

2.6. Hierarchical clustering

Once a spatial measure is obtained, stepwise clustering techniques, such as hierarchical agglomerative clustering (HAC), are a quick yet practical approach to group similar subjects without a priori knowledge of the underlying data distribution [29,30]. For example, recent studies [31,32] have used hierarchical clustering to define anatomical subgroups of patients, and test for clinical significance. Furthermore, Bruse et al. [32] investigated which distance/linkage combinations would provide the most "clinical meaningfulness" when applied to a cohort of healthy and pathological aortic arches post-surgical repair patients. Their results show that hierarchical clustering using Spearman, Correlation, or Cosine metrics [33] combined with a weighted-linkage [34] function can yield significant patient subgroups based on spatial features. We use our earlier defined spatial similarity measure, and adopt the weighted-linkage function for determining the distance between the groups when performing our hierarchical clustering.

Following a bottom-up approach where each patient was first represented as a singleton cluster, we used a hierarchical agglomerative clustering (HAC) algorithm to iteratively combine clusters in a pairwise fashion, based on the computed similarity scores and linkage distance function. Based on the results from Bruse et al.'s study [32], we chose to use the weighted-linkage function [34] when determining the distance between clusters. At each iteration, the weighted-linkage function calculates the distance between every pair of clusters, i and j, by computing the arithmetic mean of distances (i.e., similarity scores) between all points in i and j. The algorithm then combines the "nearest" (smallest distance) two clusters and continues iterating until only a single cluster remains. We report clustering results for up to six groups, because below this level some of the groups are too small to accurately assess the significance of the associations with the toxicity outcomes. Clustering was performed using the Matlab r2018a machine learning toolbox [35].

Results from hierarchical clustering are commonly summarized using a dendrogram, a tree-like structure that displays how the elements are partitioned into groups based on the computed similarity and linkage functions [36,37]. We construct such a dendrogram as described further below.



Patient 14 Tumor: Right
Similarity Score: 1.00000
Feeding Tube @ 6m: Y
Aspiration: Y

Patient 245 Tumor: BiLat.
Similarity Score: 0.86830
Feeding Tube @ 6m: N
Aspiration: N

Patient 5112 Tumor: Right
Similarity Score: 0.86830
Feeding Tube @ 6m: N
Aspiration: N

Patient 10128 Tumor: BiLat.
Similarity Score: 0.77290
Feeding Tube @ 6m: N
Aspiration: Y

Patient 84 Tumor: Left
Similarity Score: 0.76480
Feeding Tube @ 6m: N
Aspiration: N

Patient 10038 Tumor: Left
Similarity Score: 0.75770
Feeding Tube @ 6m: Y
Aspiration: N

Patient 283 Tumor: BiLat.
Similarity Score: 0.74430
Feeding Tube @ 6m: N
Aspiration: Y

Patient 10059 Tumor: Right
Similarity Score: 0.74430
Feeding Tube @ 6m: Y
Aspiration: N

Fig. 7. Example similarity ranking. Patient #14 (shown top left) is unique within the cohort, in that no other patient in the 582 patient cohort exhibits the same ten bilateral LN levels and RP involvement. Following Patient #14 are the seven closest-ranked patients (shown in left-right and top-down order) based on our spatial similarity measure. The two most similar patients share eight bilaterally involved LN levels; the next two have similar bilateral chains but either share fewer involved LN levels (Patient #10128) or possess two additional involved LN levels (Patient #84); while the last three similar patients have similar involvements but with significantly fewer LNs levels.

2.7. Statistical analysis

The patient groupings were further compared using the Rand Index [38] to determine the measure of similarity between the two measures' (spatial and nonspatial) clustering output. This measure quantifies the number of pairing agreements between two clusters into a frequency between 0.0 and 1.0, where a value of 0.0 indicates that the clusterings disagree on every pairing of samples and a value of 1.0 indicates that both clusterings are the same. Additionally, the Fisher's exact test [39] was performed on the spatial clustering to assess correlation with toxicity, as described in detail in our results. Statistical tests were performed using the Matlab 2018a statistical toolbox [35].

2.8. Visual analysis

To facilitate the assessment of our approach by clinicians, we have constructed an application to help interpret the abstracted nodal involvement of each patient in the cohort in the context of the computed similarity between patients. The visual interface (Fig. 7) consists of small multiple representations of the abstract topological map (Fig. 2 Right), and control menus which allow a specific patient to be selected and viewed. To keep the representations compact, only one side of the head and neck was abstracted; color was used to distinguish between left (green), right (purple), and bilateral (blue) involvement. The visual interface was implemented using the web technologies JavaScript, HTML, CSS, and the D3 [40] Javascript library.

To further convey the patient clustering and statistical analysis results, we created an additional informational dendrogram (Fig. 10). The dendrogram shows a tree structure that captures the paths through which smaller clusters are merged, during hierarchical clustering, into the $k = 6$ final clusters. The tree root corresponds to the entire dataset of 582 patients, whereas branches shown underneath the root correspond to the clusters formed through hierarchical clustering, and leaves at the lower-level of the tree diagram correspond to the smallest subclusters considered. The resulting six groups G1-G6 of patients ($227 + 174 + 28 + 77 + 51 + 25 = 582$ patients) are highlighted in color, along with their contributing subclusters. Toxicity statistics, along with the patient count, are displayed as minitables atop each of the six groups. An orange background in the minitable cells indicates significant correlation with specific toxicities.

Because the clustering and statistical analysis are performed over the entire dataset, and clusters often include similar, but not identical patient patterns, it is not possible to illustrate the dendrogram clusters, in the same view, with graph representations for every single patient in that cluster, in the style of the topological map visual abstraction described earlier (Fig. 2 Right). Instead, we created a thumbnail variation of the topological representation, based on the consensus nodal involvement of the patients within each subcluster (Fig. 8). This representation was derived and illustrated by considering the frequent patterns within each subcluster, as well as the less common patterns within the subcluster. First, the most frequently occurring involvement pattern for each subcluster was determined based on the consensus nodal spread in that subcluster. The consensus was determined based on a two-thirds majority involvement status (i.e., a LN level is included in the graph if 67% of the patients within that subcluster share that involvement). Next, we determined the nodes affected in less than 67% of the patients in that subcluster. We then encoded this information visually as follows: the thumbnail consensus graphs are a variation of

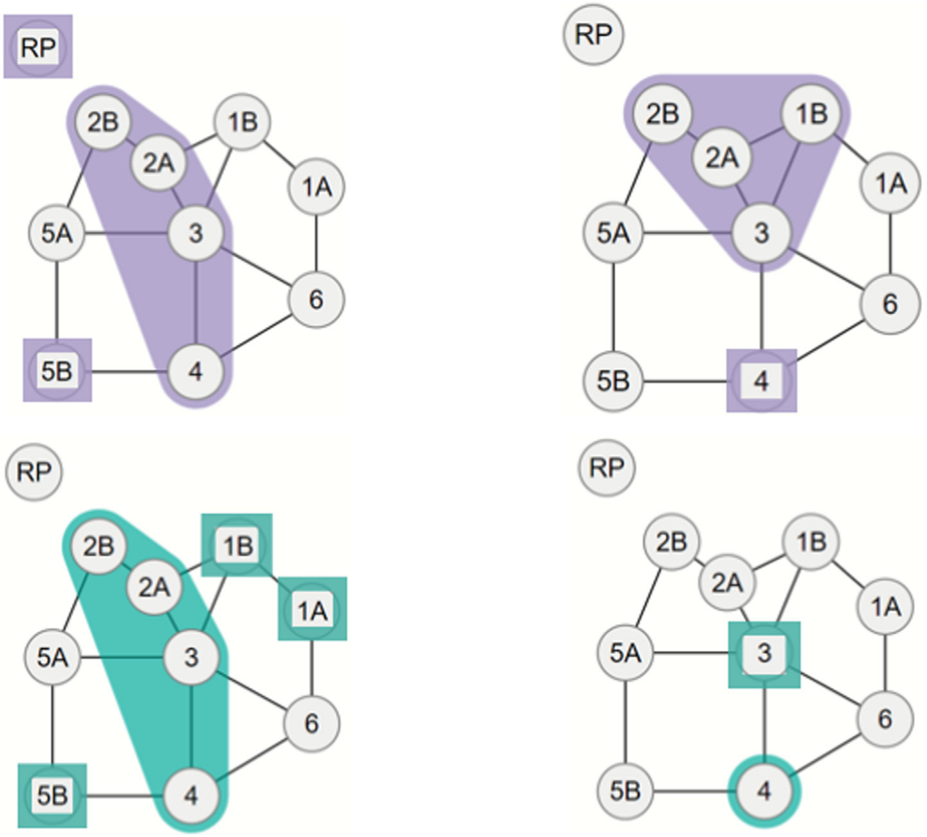


Fig. 8. Miniature consensus graph pairs for two subclusters. In this new representation, solid and outlined nodes are consensus nodes, affected in more than 67% of the patients in that subcluster. Square marks indicate nodes affected in less than 67% of the patients in that subcluster. Unilateral involvement is shown by a single consensus graph (purple), while other-side involvement (green) is indicated by an additional miniature graph stacked underneath. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the previously described graph representations (purple for the right side of the head and neck, green for the left, while accounting for symmetric matches). In this new representation, solid and outlined nodes are consensus nodes, affected in more than 67% of the patients in that cluster, while square marks indicate nodes affected in less than 67% of the patients in that cluster. Because overlapping color information is hard to read at miniature scale, unilateral involvement on only one side of the head and neck is shown by a single consensus graph, while bilateral involvement is shown by two stacked miniature graphs, one for each side of the head and neck. One such miniature graph structure is shown for each subcluster, along the X-axis of the dendrogram representation. We note that the miniature consensus graphs do not provide a complete descriptor of cluster membership.

3. Results

Because there is no other established computing methodology based on spatial information for assessing LN similarity, we demonstrate the merits of our approach using a hybrid quantitative and qualitative scheme, using real data. This scheme is further informed by the large nature of the dataset, which exceeds in scale the manual capabilities of human domain experts; and by its large number of unique and partially-matching spread patterns, which would be difficult to replicate via simulation. Under this scheme, we first demonstrate, in conjunction with clinician feedback, our method's ability to correctly discriminate patients, and we contrast this ability against a nonspatial approach. Next, we quantitatively examine the overlap between the discriminative ability of the spatial and of the nonspatial approach. We then examine quantitatively the structure of patient groupings generated by our method, and show its ability to identify groups of patients that have an increased risk of developing one or both of two known toxicities. Last, we report on the technical performance of our method.

3.1. Spatial vs. nonspatial patient discrimination

We used our spatial method to group the 582 patient dataset into six clusters ($k = 6$), as earlier described. We then used the nonspatial (categorical) approach to also group the dataset into six clusters. Since physician-based assessment was the only other existing approach for spatial LN similarity detection, a group of four radiation oncologists, all with head and neck expertise, assessed the ability of the two approaches to correctly discriminate patients. In particular, the domain experts examined in detail the pairs of patients that had been, incorrectly in their opinion, deemed similar by the nonspatial approach, and endorsed their discrimination into different clusters by the spatial approach. Similarly, they examined non-trivial sets of patients clustered together by the spatial approach, and again endorsed the result. We reproduce below and in Fig. 9 one such example of a detailed analysis.

Detailed case analysis: Fig. 9 shows a case analysis that is a representative example of the value of spatial-measure over the nonspatial measure. Shown are two patients that have drastically geometrically different LN level involvements. The spatial approach successfully discriminates between these patients (Fig. 9, bottom left). In contrast, the nonspatial approach erroneously bins together these patients (Fig. 9, bottom right). During the case analysis, the oncologists noted that Patient A possesses a bilateral involvement (gray-blue), as well as a LN level 3 unilateral involvement (green). Involvement of level 3 implies potential radiation dose to laryngeal structures and is thus a potentially meaningful correlate of radiation-associated sequelae [41]. Likewise, RP node positivity implies potential radiation dose to the superior pharyngeal constrictor muscle, which is atypical, and has the potential for specific toxicity discrimination [42]. While beyond the scope of this work, this hypothesis should be explored in a future study. In the clinicians' assessment, these are important distinctions, given prior data that shows differential swallowing toxicity as a function of superior pharyngeal constrictor versus cricopharyngeus muscles [43,44]. Furthermore, in the spatial measure, Patient A was also

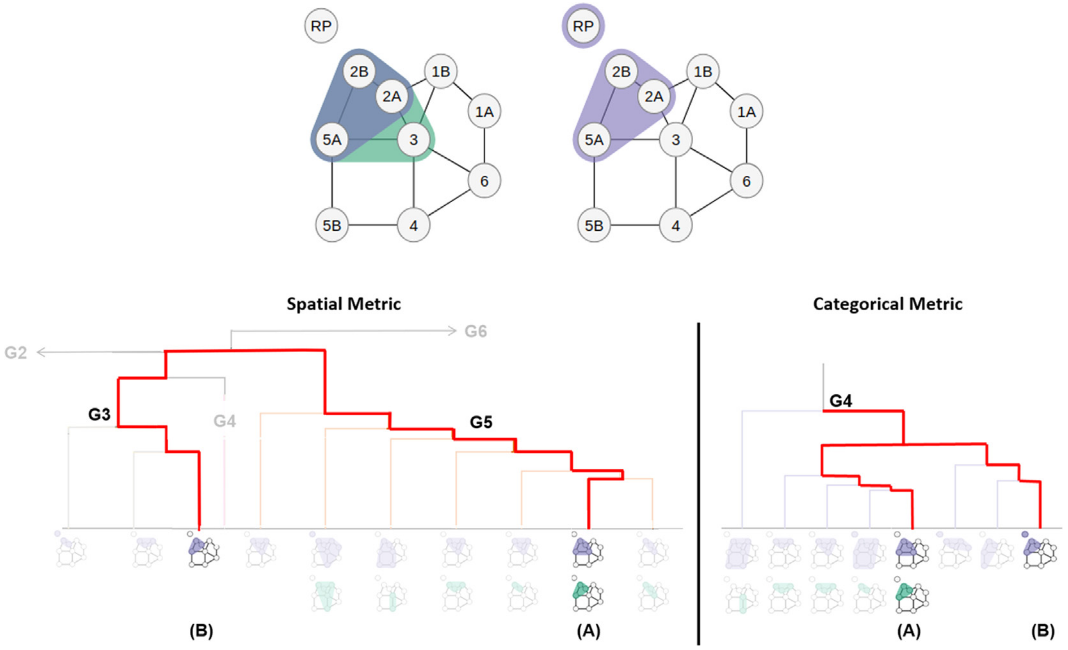


Fig. 9. Two example subjects with different groupings based on the similarity measure. Patient A (top left) possesses a bilateral nodal spread with LN level 3 involvement while Patient B (top right) only possesses a unilateral nodal spread with LN level 3 involvement. Because the spatial measure uses the geometrically different nodal involvement, it separates Patient A and B into the two main clusters, G3 and G5 (bottom left). In contrast, the nonspatial measure combines the two patients under the same main cluster, G4 (bottom right). Shown along the bottom of the X-axis are the miniature consensus graphs for the corresponding subclusters.

clustered together with other patients that have node 3 involvement, while Patient B was clustered together with no other patients that have node 3 involvement. Conversely, Patient B was primarily clustered together with patients with RP involvement (67% with RP involvement), while Patient A was not (16% with RP involvement).

During the evaluation process, the clinicians further noted that it is common practice to delineate patient groups based on bilateral involvements and the nodal spread between LN levels 2 and 3. Of the two approaches to group patients based on their lymphatic nodal spread, the clinicians indicated that the spatial similarity measure, which inherently separated patients between uni- and bilateral involvements as well as the LN level 2 and 3 nodal spread, most closely represented what is expected in a clinical setting. Due to the multiple instances of incorrect binning in the nonspatial approach, the clinicians did not recommend the further use of the nonspatial approach.

3.2. Hierarchical clustering analysis

In sum, our spatial approach was able to successfully discriminate patients based on spatial involvement in cases where the nonspatial approach failed. For example (Fig. 10 bottom rows), the spatial measure was able to discriminate between patients with significant bilateral spread and patients with no or light bilateral node involvement by placing them into separate cohorts. The spatial measure also discriminated between specific node involvement versus no involvement, regardless of pattern spread complexity. Consequently, this approach allowed for binning of patients in cohorts that were deemed by clinicians and end-users (co-authors CDF, HE, BE, AM) significantly more informative than nonspatial binning. In addition to comparing patients of the cohort, the clinicians also identified several patients whose LN levels had been previously mislabeled in the dataset due to segmentation or data processing pipeline errors.

Fig. 10 displays the informational dendrogram resulting from patient binning using the spatial measure. Miniature graphs along the bottom axis indicate the most common patterns present in each subcluster. In this dendrogram, we identified three large distinct categories by focusing on the miniature graph representations shown on the bottom. First, the branching that separates groups G2-G4 from G5, also partitions the cohort according to the degree of involvement laterality: groups G1-G3 consist of patients with no or light bilateral involvement, groups G4 and G5 of patients with significant bilateral involvement or extremely heavy unilateral involvement, and group G6 of patients with unique (singular to the cohort) nodal involvement. Next, the branching that separates groups G3 and G4, also discriminates based on LN level 3 involvement. The branching also separates the groups with no LN 3 involvement (G1, G3) from the remaining groups.

In contrast, and as discussed in detail in Section 3.1 above, the nonspatial approach failed to capture a meaningful demarcation between LN level 2 and level 3 involvement, as well as patterns of bilateral involvement. Because the nonspatial binnings were proven incorrect above (Section 3.1), whereas the miniature consensus graphs in our informational dendrogram do not provide a complete descriptor of cluster membership (as discussed in Section 2.8), constructing and reporting an equivalent informational dendrogram based on nonspatial information only is not justified.

Measure Agreement: In terms of agreement between the spatial and nonspatial approaches, we identified two identical groups between the spatial- and nonspatial-approach clusterings (G1 and G6). While these two groups represent 43% (252 patients) of the cohort, the consensus nodal involvements in each are also the simplest patterns in the cohort. For example, all 227 patients in both G1 groups possess a unilateral LN level 2 involvement, and no other node levels are involved. Similarly, the G6 patients do not have LN level 2 involvement. Furthermore, G6 groups together all the 25 unique LN level involvement patients in the cohort. Outside of these two groups, the nonspatial-approach did not have the discriminatory value of the spatial approach advocated in this paper.

After removing the two groups G1 and G6 from each of the clusterings, the computed Rand index between the spatial and the nonspatial results was a similarity measure of 55%. This value indicates that outside of the two groups G1 and G6 of simple patterns, the two approaches are significantly dissimilar in terms of how they group the patients within the cohort.

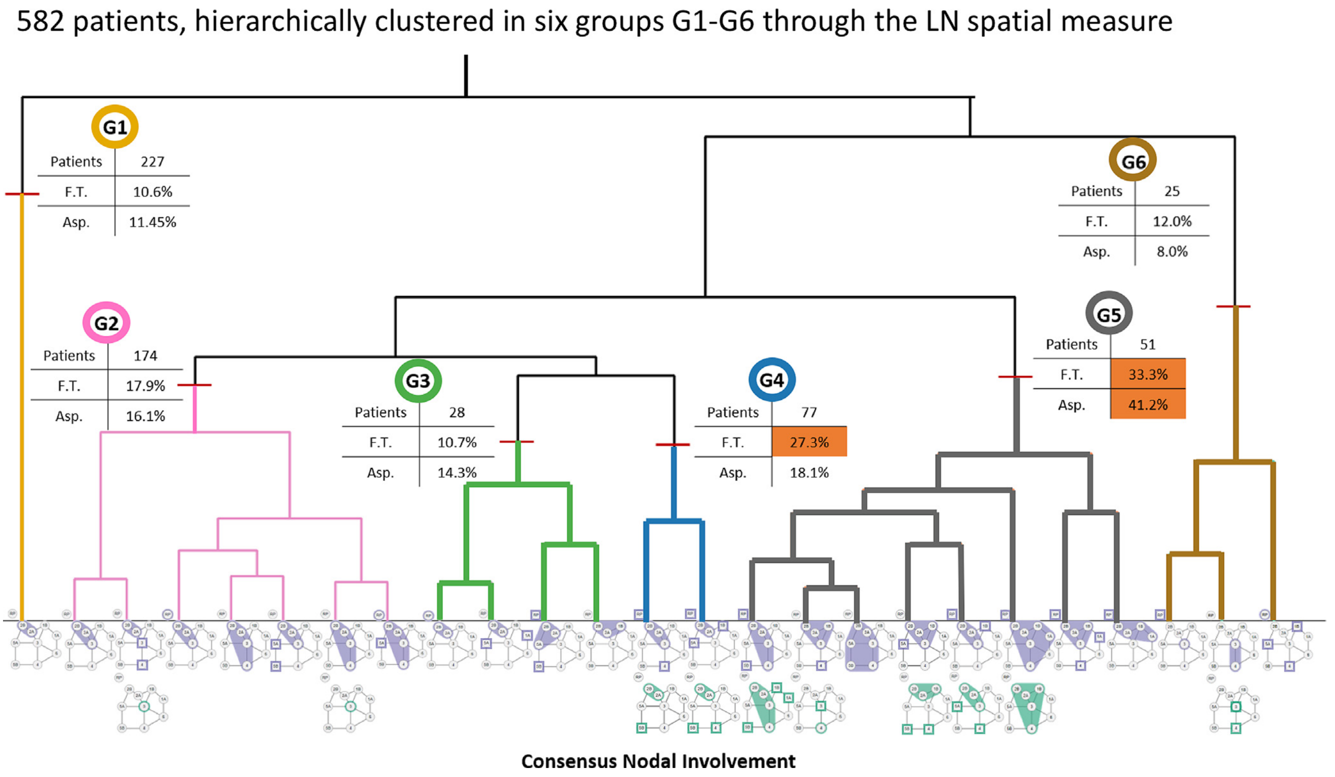


Fig. 10. Hierarchical clustering showing $k = 6$ patient spatial groups (yellow, pink, green, blue, gray, and brown), along with their toxicity correlates (mini-tables). Orange in a mini-table indicates statistically significant correlation between that group and feeding tube (F.T.) and/or aspiration (Asp) toxicities. Colored branches indicate the hierarchical clusters contributing to each of the six groups. Miniature consensus graphs along the x-axis further illustrate the sub-types present among the patients in each group. Clear distinctions are apparent, for example, between patients with bilateral vs. unilateral nodal spread (G4 and G3), or between patients with vs. without LN level 3 involvement (G4 and G5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Statistical analysis results

We use Fisher's exact test to compute and analyze the frequencies of toxicity outcomes in the different patient groups that were identified through the spatial-measure based clustering. To this end, we used two toxicity binary variables (Y/N) provided with the cohort: the post-treatment aspiration symptoms, and feeding-tube necessity at six-months. We chose the more computationally-expensive Fisher's exact test over a standard Chi-squared test because the high variation of nodal involvement patterns within the cohort yields small numbers of expected values within each group. Whereas when using Chi-squared the number of expected values for each group should be at least 5, to guarantee the significance of the p-value (otherwise a small p-value could be in fact not significant), Fisher's exact test works well on small numbers of samples.

The null hypothesis under the test is that there is no difference in the proportion of toxicity outcomes (feeding tube and aspiration rate) between the different patient groups. Statistical significance is reported assuming a level of $p < 0.01$. We note that because the nonspatial binnings are proven incorrect in Section 3.1, and because p-value testing does not capture strength of correlation, reporting an equivalent p-value analysis based on nonspatial information would be incorrect. The spatial-measure groups' p-values for feeding tube (FT) placement and aspiration rate (AR) were $p = 0.0004$ and $p = 0.0003$, respectively. The small p-values provide evidence to reject the null hypothesis, concluding that there is a statistically significant correlation between the patient groups and the toxicity outcomes.

Moreover, the odds ratio computed for the different patient groups also shows a strong association with the toxicity outcomes (Table 2). The odds ratio is defined as the ratio between the odds of having the toxicity outcome when the patient belongs to a group, and having the toxicity when the patient does not belong to the group. An odds ratio of 1 means the probability of having the toxicity is independent of the group membership. In terms of the toxicological outcomes, for FT, the spatial measure was able to identify two cohorts (G4, G5) with almost double the outcome incidence compared to the other four (G1-G3, G6). G4 and G5 had FT placement rates of 27.3% and 33.3%, respectively, while G1-G3 and G6 had rates less than or equal to 17.9%. The FT odds ratio for patients in G4 and G5 is 2.04 and 2.74, respectively. This means that patients in G4 and G5 are 2.04 and 2.74 times more likely on average to require feeding-tube at six-months. Additionally, the spatial measure identified one group (G5) with more than double the aspiration rate (41.2%) compared to the other five groups (with odds ratio 3.864). Patients in G5 are on average about 4 times more likely to develop post-treatment aspiration symptoms than other patients.

Table 2
Toxicity Outcome Distributions of the Spatial-Measure Groups.

		Feeding Tube Placement		Aspiration	
Group	*Patients*	*W/ outcome*	*% W/outcome*	*W/outcome*	*% W/outcome*
G1	227	24	10.6%	27	11.9%
G2	174	31	17.9%	28	16.1%
G3	28	3	10.7%	4	14.3%
G4	77	21	**27.3%**	14	18.1%
G5	51	17	**33.3%**	21	**41.2%**
G6	25	3	12.0%	2	8.0%
Total	582	99	17.0%	96	16.5%

Performance: We performed all computation on a 4.0 GHz Quad Core i7 machine with 32G of RAM. The average runtime to compute the similarity on the cohort of 582 patients was approximately 90 s per similarity measure. The hierarchical clustering and statistical analysis averaged 45 s to partition the patients into groups, compute the Chi-squared and Fisher's exact test, and output the statistics and dendrogram per measure.

4. Discussion

The spatial approach we introduced in this methods paper captures and ranks patients, based on their LN disease spread, correctly and more clinically accurately compared to the nonspatial (categorical) approach. Furthermore, we have shown that our novel graph-based similarity measure partitions an OPC patient cohort into clinically meaningful groups. In particular, we have shown that our spatial approach can capture groups of patients more susceptible to dysphagia toxicity (aspiration and feeding tube) based on the pattern of nodal involvement. Qualitative feedback from repeated evaluation with our collaborating clinicians further emphasized the usefulness of this approach. When presented with the informational dendrogram (Fig. 10, the most senior clinician (CDF) stated that he felt confident he could take the visualization back to his clinic that day and use it when describing the potential outcome risks alongside proposed treatment plans to his patients. Our analysis of results and the domain expert feedback support our claim that spatial correlates can provide insight into therapy strategies where treatment depends on the spatial patterns of disease, such as intensity-modulated radiation therapy for HNSCC.

We note that our results report clusters of patients that are strongly correlated with two specific outcomes. We emphasize that our method does not use the outcome information in any way to generate these clusters: the only input to our method is the LN disease spread. However, and whereas beyond the scope of this methods paper, the exploration of additional toxicities is a promising direction of future work.

In this work, we furthermore employed a Tanimoto coefficient to quantify similarity, once we constructed our feature vectors. Other coefficients may apply to variants of our problem—for example, situations that do not involve symmetry and thus do not result in non-binary vectors. We note that, whereas the exact distance coefficient used has little impact on patient rankings [24], in contrast, the spatial information we incorporate provides a mechanism to boost the similarity of connected components, and allows to differentiate patients with isolated node involvements.

In terms of limitations, our approach notes but does not explicitly incorporate into the similarity measure, the tumor location with respect to the lymph-structures (which is typically upstream in the head and neck). Other clinical applications may feature higher variability in the tumor location, and in those cases, the location of the tumor may need to be explicitly incorporated into the similarity measure. Next, we note that our evaluation was limited to one moderately sized cohort of patients. Many of these patients were referrals whose data was collected outside of the treatment facility. As a result, a significant amount of time spent working with this cohort was spent cleansing the data of malformed classifications. Furthermore, our expert feedback was limited to radiation oncology clinicians who were all members of the same clinical lab. Last but not least, our approach is constructed around a 2D graph representation that takes advantage of the symmetry about one of the principal axes of the structural model. While this approach is ideal for domains where symmetry is inherently built into the model (e.g., symmetry about the head and neck), it may also be easily extended to non-symmetric situations. In contrast, extending this approach to situations where 3D location is important would require modifications to the underlying graph representations and similarity measure.

Last but not least, while the end goal of precision medicine is personalized risk prediction and classification based on big data repositories, reaching that goal requires methods for assessing patient similarity, such as the method introduced in this paper. Our method effectively reduces complex spatial similarity to a single label, that can be used to quantify this aspect of patient similarity. We see this LN similarity method as complementary to, not competing with, radiomics similarity and genetic similarity methods [45–48]. Our similarity measure captures one of the many features that can be used in therapy response-driven decisions and predictive outcome models. While toxicity is heavily predicated on the relationship between the spatial location of involvement and the administered radiation dose, many therapy outcomes and side-effects result from other nonspatial features. A direction of future research, while beyond the scope of this work, would be to combine our spatial similarity scores with other relevant nonspatial features, such as genomics, radiomics, T-Category or patient age [49], to create a more semantically meaningful view of the patient regarding treatment response and survival. Similarly, while beyond the scope of a methods paper, an analysis of survival outcomes incorporating the spatial information, respectively the exploration of the mechanistic connections between specific toxicities and the clusters produced through this method, are important directions of future work.

5. Conclusion

In conclusion, we have introduced and evaluated a novel methodology to compare head and neck cancer patients based on their spatial patterns of LN involvement. Our approach demonstrates how the spatial location and neighborhood of the head and neck LN levels can be abstracted to a 2D topological representation, which can then be used to quantify similarity within a cohort of patients based on their extracted spatial attributes. This work also contributes two novel visual representations that provide clinicians with response-based correlates within the ranked cohort. Statistical analysis and expert feedback indicate that our spatial methodology can be useful in clinical settings. Furthermore, we show that our spatial methodology provides superior patient similarity and groupings in terms of clinical relevance when compared to the nonspatial (categorical) approach.

In an effort to make the application of this method more accessible, we also provide, in a public repository (http://github.com/uic-evl/LymphaticCancerViz/), our source code for computing and visualizing LN similarity. Whereas obviously there will be differences between different datasets and problems, we hope that knowing how to implement this method in one problem instance might help the reader more easily transfer that knowledge to another problem. The presented methodology may find application beyond the 2D head and neck lymph node analysis in other domains that feature topological structures.

Few, if any, studies have attempted to use spatial-similarity techniques to compare post-diagnosis patients and ”close the gap between mere data and useful knowledge, as desired in current Precision Medicine” [32]. Moving forward, we aim to integrate our proposed measure into a risk-prediction model. We believe that when applied to spatially-driven diseases such as OPC, approaches such as ours can play a vital role in fulfilling precision medicine's goal of maximizing the effectiveness of each patient's treatment through customized care [50].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the US National Institutes of Health, through awards NIH NCI-R01CA214825, NIH NCI-R01CA2251, and NIH NLM-R01LM012527.

References

[1] Oral Complications of Chemotherapy and Head/Neck Radiation (2018). URL

http://www.cancer.gov/about-cancer/treatment/.

[2] D. Brandizzi, M. Gandolfo, M. Lucia Velazco, R. Luis Cabrini, H. Lanfranchi, Clinical features and evolution of oral cancer: a study of 274 cases in Buenos Aires, Argentina, Medicina oral, patología oral y cirugia bucal (2008) E544–E548.

[3] B.W. Stewart, H. Greim, D. Shuker, T. Kauppinen, Defence of IARC monographs, Lancet (2003) 1300.

[4] T. Sheu, D.M. Vock, A.S. Mohamed, N. Gross, C. Mulcahy, M. Zafereo, G.B. Gunn, A.S. Garden, P. Sevak, J. Phan, et al., Conditional survival analysis of patients with locally advanced laryngeal cancer: construction of a dynamic risk model and clinical nomogram, Sci. Rep. 7 (2017) 43928.

[5] L. Zhang, A.S. Garden, J. Lo, K. Kian Ang, et al., Multiple regions-of-interest analysis of setup uncertainties for head-and-neck cancer radiotherapy, Int. J. Rad. Onc. Bio. Phys. (2006) 1559–1569.

[6] A. Nakata, K. Tateoka, K. Fujimoto, Y. Saito, et al., The reproducibility of patient setup for head and neck cancers treated with image-guided and intensity-modulated radiation therapies using thermoplastic immobilization device, Int. J. Med. Phys. Clin. Eng. Rad. Onc. (2013) 117–124.

[7] H. Elhalawani, T.A. Lin, S. Volpe, A.S. Mohamed, A.L. White, J. Zafereo, A.J. Wong, J.E. Berends, S. AboHashem, B. Williams, et al., Machine learning applications in head and neck radiation oncology: lessons from open-source Radiomics challenges, Front. Oncol. 8 (2018) 294.

[8] A. Wentzel, P. Hanula, T. Luciani, B. Elgohari, H. Elhalawani, G. Canahuate, D. Vock, C. Fuller, G.E. Marai, Cohort-based T-SSIM visual computing for radiation therapy prediction and exploration, IEEE Trans. Vis. Comp. Graph. (TVCG) 26 (1) (2019) 949–959.

[9] G.E. Marai, C. Ma, A.T. Burks, F. Pellolio, G. Canahuate, D.M. Vock, A.S. Mohamed, C.D. Fuller, Precision risk analysis of cancer therapy with interactive nomograms and survival plots, IEEE Trans. Vis. Comp. Graph. (TVCG) 25 (4) (2018) 1732–1745.

[10] J. Timar, O. Csuka, E. Remenár, G. Répássy, M. Kásler, Progression of head and neck squamous cell cancer, Cancer Metastasis Rev. (2005) 107–127.

[11] B. Cartmill, P. Cornwell, E. Ward, W. Davidson, R. Nund, et al., Emerging understanding of dosimetric factors impacting on dysphagia and nutrition following radiotherapy for oropharyngeal cancer, Head Neck (2013) 1211–1219.

[12] W. Widanagamaachchi, P. Klacansky, H. Kolla, A. Bhagatwala, J. Chen, V. Pascucci, P.T. Bremer, Tracking features in embedded surfaces: Understanding extinction in turbulent combustion, Proc. IEEE 5th Symp. on Large Data Anal. Vis. (LDAV), 2015, pp. 9–16.

[13] J. Wenskovitch, L. Harris, J. Tapia, J. Faeder, G. Marai, MOSBIE: a tool for comparison and analysis of rule-based biochemical models, BMC Bioinform. J. (2014) 1–22.

[14] C.-C. Teng, L. Shapiro, I.J. Kalet, C. Rutter, R. Nurani, Head and neck cancer patient similarity based on anatomical structural geometry, Proc. IEEE Int. Symp. Biomed. Imag. (2007) 1140–1143.

[15] B. Yener, C. Gunduz, S.H. Gultekin, The cell graphs of cancer, Bioinform. (2004) i145–i151.

[16] M.N. Gurcan, L.E. Boucheron, A. Can, A. Madabhushi, N.M. Rajpoot, B. Yener, Histopathological image analysis: a review, IEEE Rev. Biomed. Eng. (2009) 147–171.

[17] E.G.M. Petrakis, C. Faloutsos, Similarity searching in medical image databases, IEEE Trans. Knowl. Data Eng. (TKDE) (1997) 435–447.

[18] A. Kumar, J. Kim, W. Cai, M. Fulham, D. Feng, Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data, J. Digital Imag. (2013) 1025–1039.

[19] V. Gregoire, K. Ang, W. Budach, C. Grau, et al., Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines, Radiother. Oncol. (2014) 172–181.

[20] G.E. Marai, Activity-centered domain characterization for problem-driven scientific visualization, IEEE Trans. Vis. Comp. Graph. (TVCG) 24 (1) (2017) 913–922.

[21] A. Tomovic, P. Janicic, V. Kešelj, n-Gram-based classification and unsupervised hierarchical clustering of genome sequences, Comp. Methods Programs Biomed. (2006) 137–153.

[22] T.T. Tanimoto, An elementary mathematical theory of classification and prediction, Tin. Business Machines Corp., 1958.

[23] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.

[24] D. Bajusz, A. Rácz, K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminformatics (2015) 20.

[25] M. Baisantry, D.P. Shukla, Comparison of different similarity measures for selection of optimal information-centric bands of hyperspectral images, in: Observing Changing Earth, Sci. Decis. Monit., Assessment, Projection, 2017.

[26] A. Sharma, S.P. Lal, Tanimoto based similarity measure for intrusion detection system, J. Inform. Secur. 2 (4) (2011) 195.

[27] S.-S. Choi, S.-H. Cha, C.C. Tappert, A survey of binary similarity and distance measures, J. Syst. Cybernet. Informat. 8 (1) (2010) 43–48.

[28] A. Mild, T. Reutterer, Collaborative filtering methods for binary market basket data analysis, International Computer Science Conference on Active Media Technology, Springer, 2001, pp. 302–313.

[29] B. King, Step-wise clustering procedures, J Amer. Stat. Assoc. (1967) 86–101.

[30] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, Comp. J. (1983) 354–359.

[31] A. Dong, N. Honnorat, B. Gaonkar, C. Davatzikos, CHIMERA: clustering of heterogeneous disease effects via distribution matching of imaging patterns, IEEE Trans. Med. Im. (TMI) (2016) 612–621.

[32] J. Bruse, M. Zuluaga, A. Khushnood, K. Mcleod, et al., Detecting clinically meaningful shape clusters in medical image data: metrics analysis for hierarchical clustering applied to healthy and pathological aortic arches, IEEE Trans. Biomed. Eng. (TBME) (2017) 2373–2383.

[33] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Oxford Bioinform. (2000) 412–424.

[34] R.R. Sokal, C.D. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas Sci. Bull. (1958) 1409–1438.

[35] MATLAB and Statistics Toolbox Release 2018a (2018).

[36] O.Z. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer, 2010.

[37] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge Univ. Press, 2018.

[38] W. Rand, Objective criteria for the evaluation of clustering methods, J Amer. Stat. Assoc. (1971) 846–850.

[39] R.A. Fisher, Statistical Methods for Research Workers, Kalpaz, 2017.

[40] M. Bostock, V. Ogievetsky, J. Heer, D3 Data-Driven Documents, IEEE Trans. Vis. Comp. Graph. (TVCG) (2011) 2301–2309.

[41] T. Rancati, M. Schwarz, A. Allen, F. Feng, A. Popovtzer, et al., Radiation dose-volume effects in the larynx and pharynx, Int. J. Rad. Onc. Bio. Phys. (2010) S64–S69.

[42] C.R. Spencer, H.A. Gay, B.H. Haughey, et al., Eliminating radiotherapy to the contralateral retropharyngeal and high level ii lymph nodes in head and neck squamous cell carcinoma is safe and improves quality of life, Cancer (2014) 3994–4002.

[43] T. Dale, K. Hutcheson, A. Mohamed, J.S. Lewin, et al., Beyond mean pharyngeal constrictor dose for beam path toxicity in non-target swallowing muscles: Dose-volume correlates of chronic radiation-associated dysphagia (RAD) after oropharyngeal intensity modulated radiotherapy, Radiother. Oncol. (2016) 304–314.

[44] M. Kamal, A.S. Mohamed, S. Volpe, et al., Radiotherapy dose-volume parameters predict videofluoroscopy-detected dysphagia per digest after imrt for oropharyngeal cancer: Results of a prospective registry, Radiother. Oncol. (2018) 442–451.

[45] Comprehensive genomic characterization of head and neck squamous cell carcinomas, Nature 517 (7536) (2015) 576.

[46] J.C. Smith, J.M. Sheltzer, Systematic identification of mutations and copy number alterations associated with cancer patient prognosis, Elife 7 (2018) e39217.

[47] A.J. Gentles, A.M. Newman, C.L. Liu, S.V. Bratman, W. Feng, D. Kim, V.S. Nair, Y. Xu, A. Khuong, C.D. Hoang, et al., The prognostic landscape of genes and infiltrating immune cells across human cancers, Nat. Med. 21 (8) (2015) 938.

[48] Integrated genomic characterization of oesophageal carcinoma, Nature 541 (7636) (2017) 169.

[49] A.S.R. Mohamed, B.P. Hobbs, K.A. Hutcheson, et al., Dose-volume correlates of mandibular osteoradionecrosis in Oropharynx cancer patients receiving intensity-modulated radiotherapy: Results from a case-matched comparison, Radiother. Oncol. (2017) 232–239.

[50] G.H. Fernald, E. Capriotti, K.J. Karczewski, R. Daneshjou, R.B. Altman, Bioinformatics challenges for personalized medicine, Bioinform. (2011) 1741–1748.