Topological Data Analysis applied to radiomics (topiomics) data in Recurrent Prostate Cancer

Lara Cavinato[1], Lorenzo Ferrara[1], Matteo Pegoraro[2], Paola Anna Erba[3], and Francesca Ieva[1,4]

[1] MOX, Department of Mathematics, Politecnico di Milano, Milan (Italy),
[2] Department of Mathematical Science, Aalborg University, Aalborg (Denmark)
[3] Department of Medicine and Surgery, University of Milan Bicocca, Milan (Italy)
[4] Health Data Science centre, Human Technopole, Milan (Italy)

Abstract. Non-invasive imaging techniques, particularly Positron Emission Tomography (PET), have revolutionized medical diagnostics. Radiomics, an emerging field, aims to extract quantitative data from tomographic images to reveal subtle tissue alterations. Despite challenges in standardization, radiomics holds promise in linking imaging features to clinical outcomes, crucial in metastatic cancer treatment. This study explores Topological Data Analysis (TDA) to dissect intra-tumor heterogeneity in metastatic prostate cancer. Patient representation methods, including hierarchical dendrograms and persistence diagrams, are compared. Significant differences in clinical variables and treatment responses are observed across patient clusters. This research contributes to advancing precision oncology in metastatic cancer subtyping.

Keywords: topological data analysis, radiomics, prostate cancer

1 Introduction

Non-invasive imaging methods like Positron Emission Tomography (PET) have long been essential tools for medical professionals, aiding in the analysis, diagnosis, and treatment of various diseases. Radiomics, a burgeoning field, enhances traditional diagnostic techniques by extracting quantitative data from tomographic images, revealing subtleties in tissue alterations and pathological processes not visible to the naked eye. Despite technical challenges stemming from inconsistent imaging protocols and quality assurance measures, radiomics holds promise in establishing links between imaging features and clinical outcomes. Recent studies have increasingly focused on radiomic variables, recognizing tumor heterogeneity as crucial for treatment response and prognosis [1]. Addressing challenges posed by metastatic cancers, which exhibit diverse responses to therapy across multiple lesions, researchers advocate for a comprehensive approach encompassing both primary and metastatic lesions to capture the full tumor heterogeneity spectrum [2, 3]. Building on these advancements, our study seeks to delve deeper into the complexities of intra-tumor heterogeneity using advanced analytical techniques in the range of statistical methods. Specifically, we aim to leverage Topological Data Analysis (TDA) to characterize the nuanced structural patterns within tumors, transcending traditional clustering methodologies

to provide a more holistic understanding of tumor biology [4]. Building upon [5], we propose to extend the application of TDA to analyze the whole sets of lesions rather than individual ones, thereby capturing the collective heterogeneity within a tumor. TDA offers a unique perspective by considering the topological nature of data, enabling the identification of shape patterns and structural features that may hold clinical significance. Examples of TDA applications involve the use of tree-like structures, such as phylogenetic trees and hierarchical clustering dendrograms, or the use of other topological summaries such as persistence diagrams, persistence silhouettes, and persistence landscapes. In this work, we aim at comparing two different patient representation strategies, hierarchical dendrograms and persistence diagrams, in order to stratify patients affected by Recurrent, i.e., metastatic, Prostate Cancer (PCa). By embracing a multidimensional perspective, we can move closer to realizing the vision of precision oncology, where treatment strategies are tailored to the individual characteristics of each patient's tumor.

2 Methods

In this section we provide details regarding the dataset employed for the analyses and the data preoprocessing (Section 2.1), the patient representation strategies (Section 2.2), and the patient stratification pipelines (Section 2.3).

2.1 Dataset

Data from 47 Biomchemical Recurrent Prostate Cancer (BRPCa) patients who underwent PET scans using [68Ga]PSMA-11 tracer were analyzed. Patient information included age, lesion site, Gleason Score, PSA levels, number of lesions, and therapy history. PSA levels were recorded before and after Primary Treatment (TP), during hormonal therapy PET scans, and at Biochemical Recurrence (BR). While PSA is a common marker for prostate conditions, its variations can be influenced by factors beyond cancer. Evaluating disease progression based on PSA level variations relative to initial levels is important, as absolute differences may be misleading due to inherent baseline variability [6]. PET/CT image analysis yielded 129 radiomic features (specifically from the radiomic groups of GLSZM, GLCM, GLRLM, and NGTDM) per lesion, but a high rate of missing values, particularly in certain radiomic groups, was noted, possibly due to technical challenges affecting data reliability.

2.2 Patient Representation

Radiomics features were considered for patient representation and, specifically, were reduced with Principal Component Analysis (PCA) before usage. Being q the number of principal components and n_i the number of lesion belonging to patient i, we define patient i as $C_i = x_1^q, ..., x_{n_i}^q$, where x_j^q with $j = 1, ..., n_i$ are the lesions representations the radiomic space $\mathbb{R}^q$. The set is said point cloud. Two

different methods from Topological Data Analysis were employed to summarize the patients' point clouds: hierarchical dendrograms and persistence diagrams.

Hierarchical dendrograms Hierarchical clustering dendrograms have been proposed as a summary as they unveil the intrinsic relationship among points of a point cloud. A dendrogram is obtained in such a way that lesions are linked in terms of similarity relationship in their imaging characteristics. It quantifies to which extent lesions, i.e., their radiomic vectors, are similar within patients and how they get agglomerated hierarchically, one to each other. Operationally, we computed the pairwise Euclidean distances between peer lesions to generate a distance matrix, used to build a dendrogram with average linkage. The structure of the dendrogram visually represents the merging patterns within the data, reflecting the distribution of the intra-tumor variability [5].

Persistence diagrams A distinct topological method was employed, treating each patient's lesion cloud point as a Persistence Diagram (PD). In contrast to trees, PDs are part of the broader framework of Persistent Homology, a TDA technique for assessing topological data features across various spatial resolutions. This involves creating a series of topological spaces based on increasing disk radii, known as a filtration, to analyze shape characteristics at different scales. From this filtration, homology groups are extracted to describe the evolution of topological features. Dendrograms visually represent path-connected components at a fixed height, akin to clusters formed by merged points. In comparison, Persistence Diagrams condense filtration information into points $x_1, ..., x_N \in \mathbb{R}^2$ representing the appearance and disappearance of topological features, such as connected components, holes, or voids. Our focus on lesion grouping patterns led to the use of persistence diagrams, particularly highlighting connected components, where features start at a radius ρ of 0 and merge as the radius expands, symbolizing the merging of disks when their centers are within a certain distance [7]. In practical terms, the merging of two disks occurs when their centers are positioned in a way that $d(O_1, O_2) < 2\rho$.

2.3 Patient Stratification and cluster interpretation

To perform clustering we need a proper distance, which capture the different extent of tumor heterogeneity and allows the pair-wise comparison between patient-representation object, i.e., hierarchical trees and persistence diagrams.

Clustering of Hierarchical Trees After obtaining patients' trees, they were compared using an Edit Distance (d_E) metric. To compute the d_E measure between two trees, changes are made to the structure of the first tree to align it topologically with the other. These modifications include shrinking, inserting, or deleting edges, with their impact on distance computation contingent upon the length of the affected edge. This method establishes a distance metric between

trees by identifying the shortest path, requiring minimal alterations, to convert the first tree into the other. To enhance the clustering procedure's robustness, a regularization approach was implemented through a Pruning Edit distance (PTED) [5]. Utilizing this method, leaf pairs with reciprocal distances below a selected threshold ϵ are deemed sufficiently homogeneous to be merged into a single lesion. Thus, this approach contrasts with mere leaf comparison by considering distinct phenotypes identified through pruning, effectively mitigating the influence of minor differences in lesion phenotypes on the overall structure while preserving significant heterogeneity. However, due to PTED's sensitivity to the ϵ parameter, an averaging approach was employed using multiple ϵ values, each weighted according to a chosen distribution μ. This technique redefines PTED, facilitating the characterization of both pruning intensity and focus, thereby enhancing analytical flexibility and robustness. In our study, a Beta parametric model was employed to select the μ distribution, refined by comparing it with the merging heights of patients' trees to reveal the rate of small leaf merging into larger clusters and highlight variability and heterogeneity patterns within the dataset.

Clustering of Persistence Diagrams Once we acquired the persistence diagram for each patient, we established the distance between two diagrams through a matching, where each point in S is paired with a point in T or on the diagonal using a bijection $\phi : S \to T \cup (x, x) : x \in \mathbb{R}$. The distance is determined by evaluating the difference between each point and its corresponding match, calculated using a cost function. In our analysis, we utilized the q-Wasserstein distance and the Bottleneck distance, named for their associations with partial optimal transport. For $q \geq 1$, the Wasserstein metric is conventionally defined, but for $q < 1$, an extension is employed by introducing a new function $\tilde{d}_q(x, y) = (d_q(x, y))^q$, which maintains coherence with the interpretation of the classical q-Wasserstein metric. The Bottleneck distance, representing the supremum of distances and considering only the largest distance, can be viewed as the limit of the Wasserstein distance as q approaches infinity. Subsequently, the obtained distance matrices underwent the clustering algorithm, simultaneously utilizing multiple distance matrices derived from both the Bottleneck distance and the Wasserstein distance, incorporating various values of q within the range $(0, 3]$. The resulting clusters were then analyzed and characterized [8].

2.4 Cluster analysis

Following the computation of distances between patients, Hierarchical Clustering was employed, with the number of clusters set to 3 to align with the dendrogram representation. Clinical features were utilized to characterize the phenotype and treatment plan of the identified clusters. Mann-Whitney non-parametric tests were utilized for comparing distributions of continuous variables, while Chi-squared tests were employed for assessing independence of categorical variables. Pairwise one-sided comparisons between groups were conducted instead of mul-

tivariate analysis to provide a group-wise characterization, with a significance level of $\alpha = 0.1$ selected to address the impact of small sample size.

3 Results

3.1 Dendrograms-based pipeline

The Hierarchical clustering with complete linkage resulted in imbalanced cluster sizes, comprising 27, 7, and 1 elements, respectively. The last cluster was considered an outlier and subsequently discarded, while the first two were subjected to further analyses. Specifically, we examined the clusters to understand how clinical variables were influenced by and interacted with the clustering procedure. We assessed disparities between the derived groups regarding response to treatment (refractory vs. relapsing), initial PSA values and their changes over time, number of lesions, age, and Gleason score. Additionally, we evaluated the correlation between the clustering patterns and other clinical information, such as the implementation of specific therapies. In the dataset, significant differences were found for age ($p = 0.082$), number of lesions ($p < 0.001$), PSA levels variation during the PET ($p = 0.006$), and Gleason Scores, which describe the aggressiveness of the metastatic disease. Particularly, when testing the GS categorical variable, an Independence test was employed ($p = 0.06$). However, due to the limited number of observations per group, a numerical approach was also utilized, involving consideration of the numerical result C for each Score of the form of A + B = C and conducting tests based on these numerical values. This test was consistent with the previous one ($p = 0.001$); however, it's important to note that the labels A + B and B + A are not clinically equivalent. Moreover, the clusters showed significant differences in the number of lesions ($p < 0.001$), initial PSA levels ($p = 0.083$), and PSA levels variation after the TP ($p = 0.091$) and during the PET ($p = 0.006$). The clustering procedure also revealed differences in treatment therapies, with LHRHa and Abiraterone Acetate (AA) emerging as the main differentiating factors ($p = 0.06$ and $p = 0.04$, respectively).

3.2 Persistence diagrams-based pipeline

Similar to the previous section, the Persistence homology approach also categorizes the patients into three clusters. Cluster 2 contained a limited number of observations, resulting in the classification of elements within this cluster as outliers. Consequently, only clusters 0 and 1 were subjected to analysis. From a clinical perspective, the characterization of the clusters aligned with the one performed in the PTED framework. Specifically, in both datasets, significant differences were observed in the variations of PSA levels (After treatment: p = 0.04; During PET: p = 0.002) and in the number of lesions (p = 0). Similarly, a difference in the value of the Gleason score is observed only in the dataset (p = 0.05). The intuitive parallelism between the results yielded by the PTED approach and the PD approach was emphasized by directly comparing the clustering labels. A high level of concordance was evident and could be quantified through the use of the F1 score, which was 0.94.

4 Conclusions

Interpreting radiomics features extracted from PET/CT exams remains a significant challenge, and ongoing research is dedicated to establishing a robust mathematical and statistical foundation to aid treatment decision-making for patients with multiple lesions. In this study, we addressed this challenge by exploring TDA methods to leverage the intrinsic information captured by radiomic features in multi-lesion patients. Our objective was to uncover significant correlations between radiomic signatures detected in the patients under study and the distinctive characteristics observed in the phenotype and evolution of their corresponding cancers, with particular emphasis on characterizing cancer heterogeneity.

5 Acknowledgements

We acknowledge all the personnel of Medicine Department of Azienda Ospedaliero-Universitaria Pisana for the assistance during the PET/CT scans, segmentation of lesions, extraction of radiomic features and retrieval of patients' personal information from EHR. We particularly thank dr. Paola Anna Erba and dr. Costanza Bachi for their support.
L. Cavinato is funded by the National Plan for NRRP Complementary Investments - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM).
The authors acknowledge the support by MUR grant Dipartimento di Eccellenza 2023-2027.

References

1. Majumder, Shweta, et al. "State of the art: Radiomics and Radiomics related Artificial intelligence on the road to clinical translation." BJR— Open (2023): tzad004.
2. Chang, Enoch, et al. "Comparison of radiomic feature aggregation methods for patients with multiple tumors." Scientific Reports 11.1 (2021): 9758.
3. Cavinato, Lara, et al. "Radiomics-Based Inter-Lesion Relation Network to Describe [18F] FMCH PET/CT Imaging Phenotypes in Prostate Cancer." Cancers 15.3 (2023): 823.
4. Chazal, Frédéric, and Bertrand Michel. "An introduction to topological data analysis: fundamental and practical aspects for data scientists." Frontiers in artificial intelligence 4 (2021): 108.
5. Cavinato, Lara, et al. "Imaging-based representation and stratification of intra-tumor heterogeneity via tree-edit distance." Scientific reports 12.1 (2022): 19607.
6. Simon, Nicholas I., et al. "Best approaches and updates for prostate cancer biochemical recurrence." American Society of Clinical Oncology Educational Book 42 (2022): 352-359.
7. Divol, Vincent, and Théo Lacombe. "Understanding the topology and the geometry of the space of persistence diagrams via optimal partial transport." Journal of Applied and Computational Topology 5 (2021): 1-53.
8. Givens, Clark R., and Rae Michael Shortt. "A class of Wasserstein metrics for probability distributions." Michigan Mathematical Journal 31.2 (1984): 231-240.