Received: 21 February 2021 | Revised: 9 September 2021 | Accepted: 13 September 2021

DOI: 10.1002/mp.15255

RESEARCH ARTICLE

MEDICAL PHYSICS

Persistent homology of tumor CT scans is associated with survival in lung cancer

Eashwar Somasundaram[1] | **Adam Litzler**[2] | **Raoul Wadhwa**[3] | **Steph Owen**[4] | **Jacob Scott**[1,3,4,5]

[1] Case Western Reserve University School of Medicine, Cleveland, Ohio, USA

[2] University of Colorado Boulder, Department of Applied Mathematics, Boulder, Colorado, USA

[3] Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

[4] Lerner Research Institute, Department of Translational Hematology and Oncology Research, Cleveland, Ohio, USA

[5] Taussig Cancer Institute, Department of Radiation Oncology, Cleveland, Ohio, USA

Correspondence
Eashwar Somasundaram, Case Western Reserve University School of Medicine, 9501 Euclid Ave, Cleveland, OH 44106, USA.
Email: eashwarv.soma@gmail.com

Funding information
Case Comprehensive Cancer Center; National Cancer Institute of the National Institutes of Health, Grant/Award Number: R37 CA244613

Abstract

Purpose: Radiomics, the objective study of nonvisual features in clinical imaging, has been useful in informing decisions in clinical oncology. However, radiomics currently lacks the ability to characterize the overall topological structure of the data. This niche can be filled by persistent homology, a form of topological data analysis that analyzes high-level structure. We hypothesized that persistent homology features quantified using cubical complexes could be extracted from lung tumor scans and related to survival.

Methods: We obtained segmented computed tomography (CT) lung scans (n = 565) from the NSCLC-Radiomics and NSCLC-Radiogenomics datasets in The Cancer Imaging Archive. These scans are three-dimensional images whose pixel intensity corresponds to a number of Hounsfield units. Cubical complexes are a topological image analysis method that effectively analyzes the number of topological features in an image as the image is thresholded at different intensities. We calculated a novel output called a feature curve by plotting the number of zero-dimensional (0D) topological features counted from the cubical complex filtration against each Hounsfield value. This curve's first moment of distribution was utilized as a summary statistic to show association with survival in a Cox proportional hazards model. We hypothesized that persistent homology features quantified using cubical complexes could be extracted from lung tumor scans and related to survival.

Results: After controlling for tumor image size, age, and stage, the first moment of the 0D topological feature curve was associated with poorer survival (HR = 1.118; 95% CI = 1.026–1.218; p = 0.01). The patients in our study with the lowest first moment scores had significantly better survival (1238 days; 95% CI = 936–1599) compared to the patients with the highest first moment scores (429 days; 95% CI = 326–601; p = 0.0015).

Conclusions: We have shown that persistent homology can generate useful clinical correlates from tumor CT scans. Our 0D topological feature curve statistic predicts survival in lung cancer patients. This novel statistic may be used in tandem with standard radiomics variables to better inform clinical oncology decisions.

KEYWORDS
computed tomography, lung Cancer, radiomics, persistent homology, topological data analysis, tumor Imaging

1 | INTRODUCTION

The use of radiomics in tumor imaging has been an increasingly popular area of research to better inform cancer diagnosis, treatment, and prognosis.[1] Geometric properties of tumors such as size, surface area, and volume are commonly studied features among more non-visual features such as texture.[2] We propose extending the utility of radiomics by studying topological properties. In contrast to the local structural focus of geometry, topology focuses more on global structure.

As topological properties would focus on general tumor shape properties, they are theorized to be more robust to noise and could capture information not found in traditional radiomic features. Intuitively, one can describe several topological differences between malignant and benign tumors. Benign tumors are well connected and have a homogeneous shape, whereas malignant tumors are more likely to have diffuse spread and necrotic cavities.[3] Quantifying the number of topological features may be useful in predicting patient survival. In fact, pathology already uses shape properties in the context of Gleason scores, which is a measure of prostate cancer severity based on prostate gland shape.[4]

Persistent homology is a popular technique in the -omics sciences to describe the topological features of large data sets. Persistent homology has already been used in cancer biology from genomics to histology.[5,6] A statistical measure inspired by persistent homology called the smooth Euler characteristic has been developed to predict clinical outcomes from glioblastoma tumor imaging.[7] Persistence images, an alternative representation of persistent homology, have been used in machine learning models to classify MRI images of hepatic tumors.[8] Although radiomics and persistent homology have largely existed in separate worlds, they share similar challenges. In radiomics, one challenge is to find the most informative image features for analysis.[9] In persistent homology, a similar challenge is finding the most useful topological data representation for visualization, statistical comparison, and predictive modeling. We were interested in whether image features related to topology in lung CT scans were associated with survival. We thought that this might be a reasonable connection due to the success of similar aforementioned methodology in glioblastoma and hepatic tumors. We studied non-small-cell lung cancer (NSCLC) due to the large data availability in The Cancer Imaging Archive. In addition, prior work using this data has established that traditional radiomics is associated with patient outcomes using these CT scans.[10] This allowed us to check whether our new metric is also associated with clinical outcomes in a data set proven to be amenable to traditional radiomics.

As we hypothesized earlier that an increased number of topological features may correlate to a more malignant tumor, we wanted to create a topological data representation that captured the quantity of topological features of a CT tumor scan. In this project, we create a new radiomics variable using the statistical summary variables of the zero-dimensional (0D) topological feature curve, our method of representing topological feature quantity. We describe this further in Section 2. To our knowledge, no work has been done in using persistent homology to characterize survival in cancer patients using such a summary statistic. We show through a discrete and continuous analysis that moment 1 of our 0D topological feature curve (i.e., average number of 0D features) is significantly associated with worse survival. NSCLC accounts for 85% of lung cancer in the United States and carries a dismal 5-year survival rate of 15%.[11] Developing models that may 1 day be able to predict patient outcome is of great interest to providers and patients alike.

2 | PERSISTENT HOMOLOGY BACKGROUND

Topological data analysis (TDA) encompasses a broad set of techniques used to highlight topological patterns in data. Computing persistent homology using cubical complexes is a popular method within TDA for describing topology in imaging data.[12] Counting topological features on an image require a binary black and white image, so we cannot count the topological features from the grayscale computed tomography (CT) scan directly. However, we can convert the CT scan into a series of binary black and white images (i.e., filtration) from which the topological features can be counted. In the context of persistent homology, cubical complexes are essentially synonymous with these binary images. A full mathematical description of cubical complexes and similar methodology are described in "Topological Data Analysis" in Proceedings of Symposia in Applied Mathematics.[13]

For each CT scan, we create a series of binary images (i.e., cubical complexes), one for each Hounsfield unit (HU) filtration value. For example, if the filtration value is −900 HUs, then all pixels with units less than or equal to −900 are colored black, and all pixels above are colored white. Topological features can be counted from these binary images. A particular topological feature can be described by dimension and the range of HU filtration values across which the feature is found.

We show this filtration process with slice 13 of scan 1 in Figure 1(a). This slice is shown as it demonstrates a potential connection between tissue characteristics (calcifications) and topological features. At low HU thresholds, the image is mostly white. Islands of black pixels are considered connected components or 0D features. Tissue necrosis may result in a more fragmented tumor, which would increase the number of 0D features. "Lakes" of white pixels are considered holes or 1D

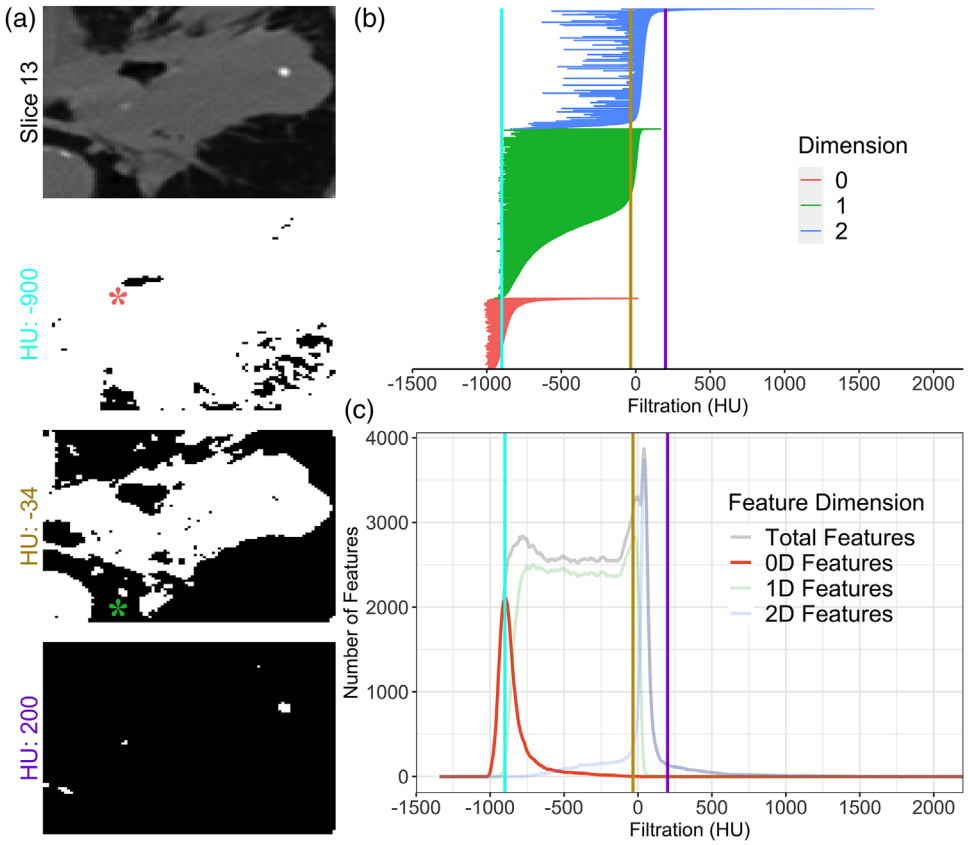

FIGURE 1 Methodology to generate cubical complexes. (a) An example slice from the Radiomics dataset. Each pixel value in a CT scan is described by a Hounsfield unit, which typically range from −1024 to 3071. A cubical complex is created by selecting a Hounsfield value as a filter. Any pixel below or equal to this filter is colored black, and any pixel above is colored white, creating a binary image that becomes the cubical complex. Topological features can then be counted from each binary image. An example of a 0D feature is indicated by the red asterisk. An example of a 1D feature is indicated by the green asterisk. Two-dimensional features are not shown as they only appear when considering the cubical complexes of all the slices together as a 3D structure. (b) We plot a barcode diagram representing the persistent homology of each topological feature. Color represents dimension, and the range represents the Hounsfield filtration unit range during which the feature existed. The colored vertical lines represent the Hounsfield units from the filtration values shown in the binary images. (c) We create three topological feature curves by summing the number of topological features by dimension at each Hounsfield unit filtration value. A fourth topological feature curve is generated by summing the total number of topological features regardless of dimension. The 0D topological feature curve is highlighted as we specifically use the moments of distribution of this curve for our survival analysis

features. Two-dimensional features are not shown in Figure 1 as they only appear when considering the cubical complexes of all the slices stacked together as a three-dimensional structure. Intuitively, they can be thought of as volumetrically contiguous groups of white pixels existing in interior of the tumor image surrounded by black pixels. Calcifications would likely manifest as 1D or 2D features late in the filtration, which is shown when the HU threshold is 200 in Figure 1(a). Figure 1 shows the pipeline being applied to a single slice for demonstrating intuition, but in this project, our pipeline is applied to the full 3D image array.

Topological barcodes are one of the most popular ways to visualize persistent homology.[14] Each topological feature in a barcode diagram is given a color to represent its dimension and a horizontal bar that spans the HU filtration values where the feature can be found. The barcode diagram in Figure 1(b) represents the entire cubical complex persistent homology of scan 1. As we were interested in topological feature quantity, we developed an alternative representation of persistent homology called the topological feature curve.

The topological feature curve is a transformation of the barcode diagram that counts the number of bars (i.e., topological features) at each Hounsfield filtration value. The number of bars is then plotted against each HU filtration value. In these curves, the full range of HU filtrations are used (−1024 to 3071). We can construct four topological feature curves, one for each feature dimension, plus a fourth curve that counts all features regardless of dimension. The topological feature curves in Figure 1(c) represent all feature curves in scan 1.

We used the raw moments of the distribution of the resulting 0D topological feature curve as predictor variables for our survival analysis. The *nth* raw moment is expected value of X^n where X represents our

feature curve distribution. The exact mathematical calculation is described in Supporting Information Table S2. We focused on the 0D curve's characteristics because it showed the most variability across tumor scans. Incorporating all of the curves moments (16 potential variables) in our survival model would not have been feasible given our sample size. We believed that the first four moments would incorporate sufficient information about the topological feature curve while avoiding the conflict of adding too many variables in our Cox proportional hazards model. Similar topological summary statistics have been described in analyzing brain connectomes of individuals with Attention-deficit/hyperactivity disorder using 0D Vietoris–Rips complex persistent homology features.[15] Likewise, we also focus our analysis on 0D features; however, we generate our topological features using cubical complexes.

3 | METHODS

The goal of this pipeline is to determine whether calculated variables describing the topology of a tumor are associated with survival in a Cox proportional hazards model. Four topological variables representing 0D features or connectedness of a tumor's CT scan were used in this Cox model. Derivation of these variables is further described in Section 2.

All images and segmentation objects were obtained from The Cancer Imaging Archive.[16] Within TCIA, the NSCLC-Radiomics (n = 422) and NSCLC Radiogenomics (n = 211) cohorts provided CT scans and clinical data.[10,17–19] Of the 633 total scans, 565 also had segmentation data, which was necessary for our pipeline. Figure 2 shows the exclusion and inclusion criteria for each analysis.

Figure 3 shows an overview of the publicly available data pipeline using R (v3.6.1) and Python (v3.7.6).[20–22] Within R, a cubical region that encapsulated the primary tumor was delineated using the segmentation file coordinates of the gross tumor volume, which had been manually delineated by radiation oncologists from the original datasets. This data object was exported to Python to compute cubical complex persistence homology (functionality in R not present at time of analysis). Within R, the raw persistence homology from Python was transformed into topological feature curves, which plot the topological features against each Hounsfield filtration value. The first four moments of the 0D topological feature curve were used as predictor variables alongside clinical covariates in our survival analysis.

For the survival analysis, univariate and multivariate Cox proportional hazard modeling was performed for the moments, tumor image size, cancer stage, age, and sex on the combined Radiomics and Radiogenomics cohort. These were all the clinical variables

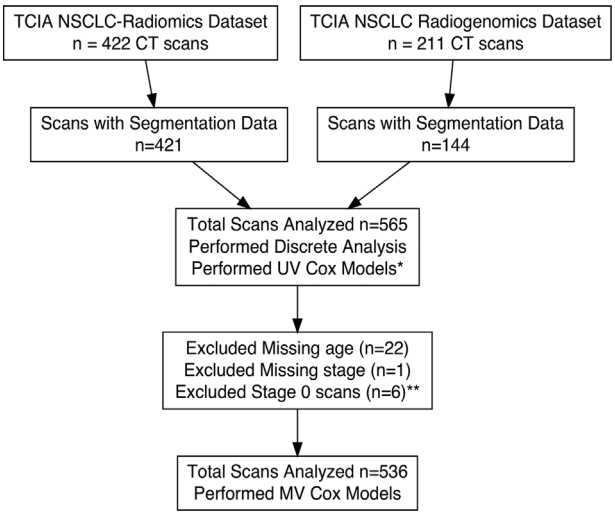

FIGURE 2 Flowchart of tumor scan inclusion and exclusion. A total of 565 tumors had both segmentation data and CT scans allowing for computation of cubical complexes in our pipeline. *The Univariate Cox models excluded patients with missing variables. **The multivariate Cox model excluded six additional patients who were stage 0 because the models did not converge (Supporting Information Table S4)

common to both the Radiomics and Radiogenomics datasets. Increasing the number pixels of an image likely increases the topological feature count without necessarily changing the overall topological structure of the tumor image. For this reason, we included tumor image size (pixels per slice × number of slices) as a covariate in the multivariable Cox Hazard model to control for potential confounding with our 0D topological feature curve. All survival times were directly provided in the NSCLC-Radiomics data set. This survival time was not directly provided in the NSCLC-Radiogenomics data set, but the date of imaging and date of death or last follow-up were provided. Survival times in the NSCLC-Radiogenomics were calculated by subtracting the date of death/last follow-up from date of imaging. The stage 0 predictor variable did not converge because none of the six patients with stage 0 cancer died during the study, so these patients were excluded (Supporting Information Table S4).

Log-rank test was used to compare all four moment 1 quartiles in the Kaplan–Meier curves, and a post-hoc log-rank test was used to make pairwise comparisons. Unpaired t-test statistics (for normally distributed data) and Wilcoxon rank sum tests (for nonnormally distributed data) were used to descriptively compare patient characteristics between the Radiomics and Radiogenomics data sets. Chi-squared test was used to compare categorical variables.

An open-source, reproducible pipeline is available at https://github.com/eashwarsoma/TDA-Lung-Phom-Reproducible with detailed instructions. All packages and libraries used were open source. In R, we used `oro.dicom` v0.5.3, `oro.nifti` v0.10.3, and

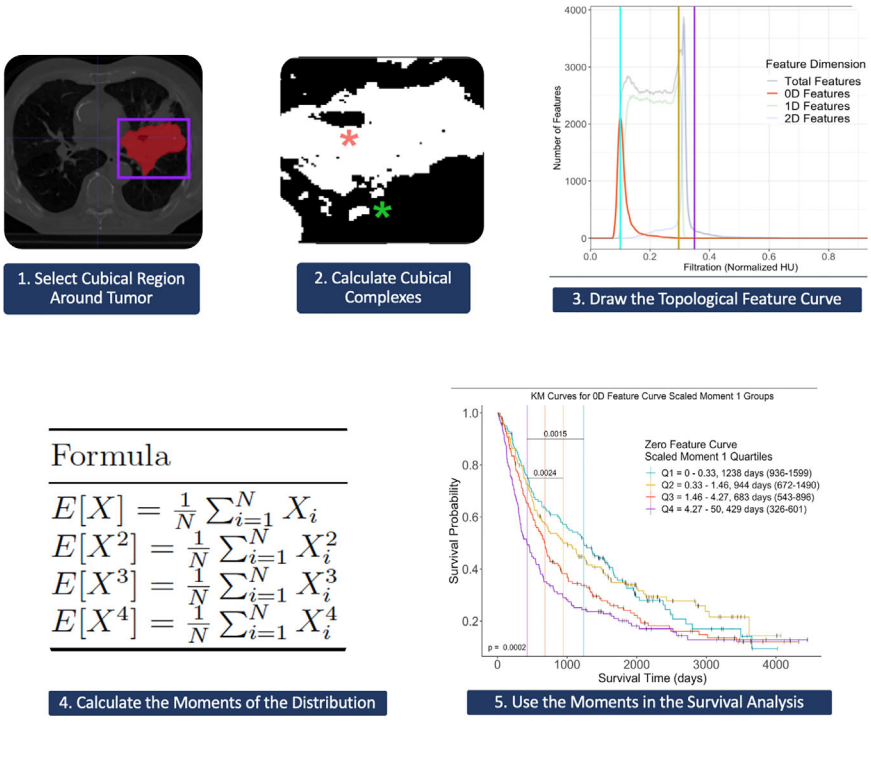

FIGURE 3 Overview of data pipeline. (1) Using the tumor segmentation data's coordinates, we computed the persistent homology. (2) Persistent homology was computed using cubical complexes and (3) transformed into a feature curve as described in Figure 1. (4) The raw moments of the 0D feature curve were calculated and (5) used as predictor variables in our survival analysis (KM Curves and Cox Proportional Hazards). The pink and green asterisks in the second panel point to an example of a 0D and 1D feature in the filtered image

`RNifti` v1.1.0 to read and process the CT DICOM scans and NIfTI segmentation objects[23–25]; `dplyr` v1.0.0, `plyr` v1.8.6, and `reshape` v0.8.8, for data wrangling[26–28]; `reticulate` v1.1.6 to convert R data structures to compatible .npy files for the Python scripts[29]; `survival` v3.2-3, `rstatix` v0.6.0, and `survminer` v0.4.7 for survival analysis and statistics[30–32]; `tableone` v0.11.2 to generate table one data[33]; and `ggplot2` v3.3.2, `TDAstats` v0.4.1, `grid` v3.6.3, `png` v0.1-7, `ggplotify` v0.0.5, `ggpubr` v0.4.0, `gt` v0.2.2, `paletteer` v1.2.0, `DiagrammeR` v1.0.6.1, `DiagrammeRsvg` v0.1, `ggfortify` v0.4.10, `rsvg` v2.1, and `gridExtra` v2.3 for data visualization and export.[34–46]

In Python, we used the built in `glob` and `pathlib` libraries to recursively read in the file objects produced from the R code; `numpy` v1.18.5 to create data structures that represented the processed tumor scans numpy objects from R[47]; `gudhi` v3.0.0 to compute the persistent homology using cubical complexes[48]; and `csv` v1.0 library to output the computed persistent homology as csv files to be processed again in R.

We used the dcmqi open source Bash library[49] to convert segmentation files into a more compatible NIfTI file format. We used ITK-SNAP v3.8.0 as the DICOM viewer to visually ensure our cubical segmentation properly highlighted the tumor region of interest.[50]

4 | RESULTS

Table 1 gives a descriptive overview of the combined two lung scan data sets. These data sets were obtained from The Cancer Imaging Archive.[16]

TABLE 1 Descriptive table of the patient scans used in this study. These data represent the characteristics of the patients pooled together from the NSCLC-Radiogenomics and NSCLC-Radiomics databases. All patient scans were downloaded from The Cancer Imaging Archive.[16] The moments of the 0D feature curves are presented as median [IQR] because the data were not normally distributed. Tumor image size and follow-up time were also not normally distributed and also presented as median [IQR]. All other data are presented as mean (SD) or as a number (proportion)

	Combined cohorts	Proportion Missing
Total sample size	559	
Vital status (% Dead)	424 (75.8)	0.0
Moment 1	24.23 [5.57, 70.40]	0.0
Moment 2	9.49×10^3 [7.72×10^2, 5.99×10^4]	0.0
Moment 3	3.92×10^6 [1.31×10^5, 6.56×10^7]	0.0
Moment 4	1.97×10^9 [2.25×10^7, 8.76×10^{10}]	0.0
Tumor image size	4.60×10^4 [1.19×10^4, 1.30×10^5]	0.0
Age	68.36 (9.78)	3.9
Stage		0.2
Stage I	180 (32.3)	
Stage II	66 (11.8)	
Stage IIIa	131 (23.5)	
Stage IIIb	177 (31.7)	
Stage IV	4 (0.7)	
Sex (% Female)	166 (29.7)	0.0
Follow-up for alive patients (days)	1975.00 [1440.50, 2984.50]	0.0

TABLE 2 Cox proportional hazard model. Both univariate and multivariate Cox hazard models show that moment 1 of the 0D feature curve is associated with poorer survival. Data are shown as hazard ratio (95% confidence interval). Bolded predictor variables were significant in both the univariate and multivariate models. Italicized predictor variables were only significant in the univariate model. The moments of distribution and tumor image size were rescaled to lie between 0 and 50 units. Linear scaling does not alter HR significance but does change the magnitude of the HR and CI to a more interpretable value. Patients with stage 0 tumors were removed from this model as the stage 0 versus stage I variable did not converge in the model (Supporting Information Table S4)

	Univariate model HR	*p*-Value	Multivariate model HR	*p*-Value
Age	1.018 (1.007–1.029)	0.0019	1.025 (1.013–1.037)	4.0×10^{-5}
Male versus Female	1.229 (0.994–1.52)	0.057	1.128 (0.898–1.417)	0.30
Scaled Moment 1	1.019 (1.005–1.033)	0.0082	1.118 (1.026–1.218)	0.011
Scaled Moment 2	1.004 (0.982–1.026)	0.74	0.766 (0.53–1.106)	0.16
Scaled Moment 3	0.992 (0.94–1.047)	0.77	1.282 (0.412–3.991)	0.67
Scaled Moment 4	0.985 (0.893–1.088)	0.77	0.995 (0.369–2.681)	0.99
Scaled Tumor Image Size	1.015 (1.003–1.026)	0.01	1.014 (0.983–1.046)	0.38
Stage II versus I	1.249 (0.884–1.765)	0.21	1.028 (0.71–1.489)	0.88
Stage IIIa versus I	1.695 (1.304–2.203)	8.0×10^{-5}	1.742 (1.317–2.305)	0.0001
Stage IIIb versus I	1.626 (1.272–2.077)	0.0001	1.483 (1.127–1.951)	0.0049
Stage IV versus I	1.279 (0.406–4.028)	0.67	1.825 (0.572–5.818)	0.31

Overall, the NSCLC Radiomics and NSCLC Radiogenomics dataset had enough similar population features to justify combining them into a single cohort. The image topological features and age were not significantly different between both cohorts. The mean follow-up time for the censured patients (alive at the end of the study) was higher in the Radiomics cohort; though both feature relatively long median follow-up times of at least 5 years. Supporting Information Table S1 more exactly compares the differences between the two cohorts.

We first performed a discrete analysis comparing survival quartiles by their moments of distribution of the 0D feature curves. This was performed primarily for intuition showing what the median topological feature curves looked like between survival groups. Supporting Information Figures S1 and S2 and Table S3 show and quantify this difference.

The discrete analysis is not as rigorous as a Cox model, which would measure the impact of the moments of the 0D topological feature curve on survival outcomes. We verified that the assumptions of proportional hazards were met through visual appraisal of the Schoenfeld residuals plot, which is shown in Supporting Information Figure S4. Table 2 shows the results of our Cox proportional hazard model.

Increasing age had a significant effect on survival (multivariate HR: 1.025; 95% CI: 1.013–1.037; $z = 4.109$; $p < 0.001$). Sex did not have a significant effect on survival (multivariate HR: 1.128; 95% CI: 0.898–1.417; $z = 1.037$; $p = 0.30$). Increasing stage also had a significant effect on survival except for stage IV, which is likely due to small sample size ($n = 4$). Moment 1 of the 0D topological feature curves had a significant effect on survival (multivariate HR: 1.118; 95% CI: 1.026–1.218; $z = 2.547$; $p = 0.011$). Moment 1 of the 0D topological feature curve significantly predicted survival even after controlling for tumor image size. Tumor image size only had a significant effect on survival in the univariate model (univariate HR: 1.015; 95% CI: 1.003–1.026; $z = 2.565$; $p = 0.01$). Figure 4 shows a forest plot of the Cox model. The non-significant Stage IV versus Stage I predictor variable was removed as its wide confidence interval interfered with data visualization.

We used Kaplan–Meier curves (Figure 5) to visually justify our survival results. Patients were divided into four groups based on scaled moment 1 quartiles. The median survival is depicted by the colored vertical lines and stated with 95% CI in the legend. The four groups had significantly different survival distributions by the log-rank test ($\chi^2 = 19.7, p < 0.001$). The median survival of the first quartile (1357 days; 95% CI: 1028–1661) was significantly different from the median survival of the fourth quartile (429 days; 95% CI: 326–601; $\chi^2 = 13.4$; $p = 0.0015$). The median survival of the second quartile (944 days; 95% CI: 672–1490) was also significantly different from the median survival of the fourth quartile ($\chi^2 = 12.5$; $p = 0.0024$).

5 | DISCUSSION

As the "-omics" sciences continue to become more present in the clinic, we need stronger analysis techniques that make meaning out of the massive amount of data. Compared to the other "-omics," radiomics may be the most useful from a clinical perspective. Obtaining repeated tumor CT and MRI scans is easier for a hospital and less invasive for a patient compared to repeated tumor sequencing (genomics) or mass spectrometry (proteomics). Therefore, it would be quite valuable if

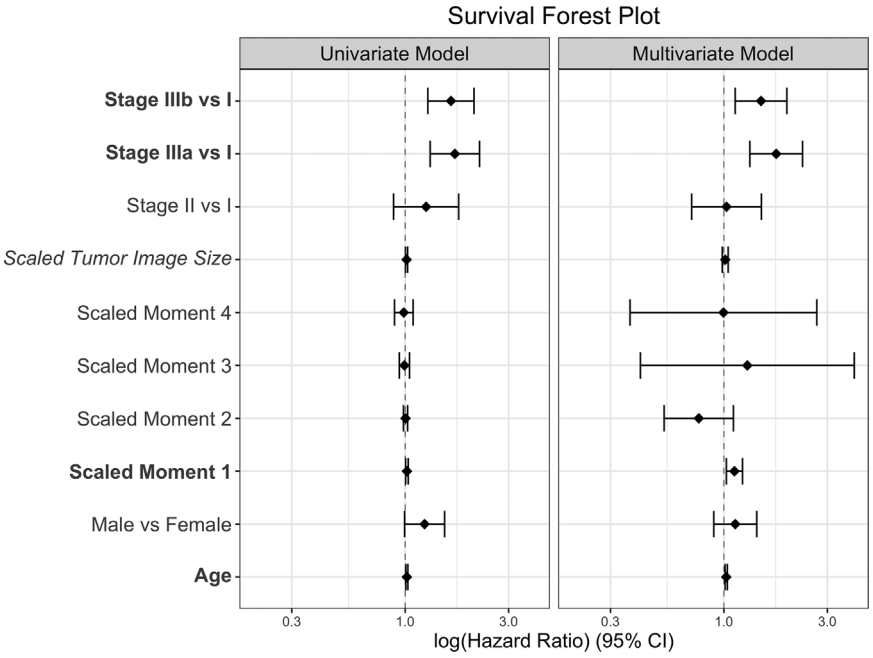

FIGURE 4 Survival forest plot shows moment 1 predicts poorer survival even after controlling for known tumor prognosticators such as stage. The horizontal axis is log transformed for better data visualization. Span of each data point represents the 95% CI. The Stage IV versus Stage I data point had a very wide CI, so it was removed from this plot for data visualization. The Stage IV versus Stage I Hazard Ratio is shown in Table 2. Bolded predictor variables are significant in both the univariate and multivariate models. Italicized predictor variables are only significant in the univariate model

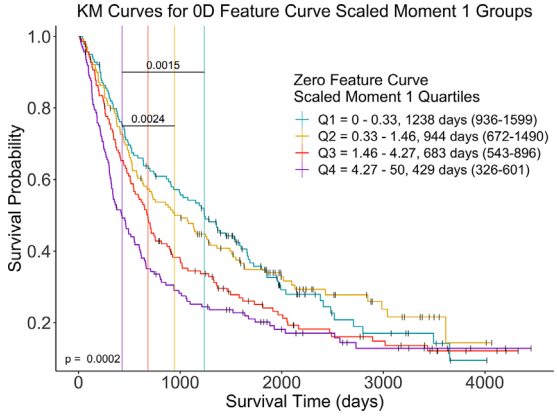

FIGURE 5 Patients with larger moment 1 values of the 0D feature curve had poorer survival. Patients were divided into quartiles based on the moment 1 value of the 0D topological feature curve. Q1 through Q4 represent Quartile 1 through Quartile 4. The exact quartile values and median survival of each quartile are shown in the figure legend. On each survival curve, the black vertical line represents a right censored event or the survival time of a patient who was still alive at last follow-up. Vertical colored lines represent median survival. The result of the log-rank comparing all survival curves are shown in the bottom left. Post-hoc log-rank analysis with Bonferroni correction was performed on survival curve pairs. The significant results are shown as horizontal lines connecting the median survival of the groups with significant differences

persistent homology could uncover additional meaning from imaging data. We show that even after controlling for tumor stage, age, sex, and image size, the first moment of the 0D topological feature curve is a significant predictor of survival. This effect was significant across two independent data sets indicating a resilience of our methodology to batch effect. We considered tumor image size as a confounding possibility because the number of topological features would be expected to increase with image size. However, the effect of moment 1 of the 0D feature curve on survival remained significant in the multivariate model, and about 30% of the variance in moment 1 of the 0D topological feature curve was not explainable by tumor image size alone (Supporting Information Figure S5).

It is hard to describe biological and medical correlations of the 0D feature curve at this point. Images with a greater number of 0D features would have a more heterogenous distribution of gray-scale values compared to images with a fewer number of 0D features. We conjecture that increased 0D features correlate with a tumor that is heterogeneous in tissue composition because different tissue types result in different gray-scale values in a CT scan. In addition, a patchy disconnected tumor may also be reflective of an increased number of 0D features as disconnected foci of gray-scale values would result in more features even if the whole tumor volume is the same compared to a well-localized tumor. We emphasize that these connections are conjectures at this point. To truly understand the biological meaning of the topological feature curve, we would likely need an exhaustive histopathological analysis to draw any definitive conclusions. Modern day statistical tools and data analysis techniques allow us to draw useful inferences about patient outcome without necessarily having to understand the underlying biology.

Despite these promising results, there remain limitations in our study. Though the pipeline worked on a combined set of tumor scans from two separate research studies, the imaging parameters across the two sets were similar. It remains unknown how sensitive this technique is to imaging parameters such as CT machine model, which has been shown to affect calculation of standard radiomic features.[51] Although we controlled

for some important clinical variables as potential confounders, there are other relevant clinical variables such as EGFR gene status we were unable to test due to lack of data. We could only control for relevant clinical variables common to both data sets. Factors such Karnofsky Performance Score (KPS) and nodal status are known to be predictive to overall survival in lung cancer.[11] Ideally, we would have included these variables had they been available to assess how our statistic interacts with them in survival models. However, we believe that we have sufficient evidence from our models to say that our persistent homology metric is correlated with overall survival. For our current study, we know all of the analyzed scans were for planning radiotherapy or surgery. However, all of the patients' prior clinical history is a black box, which is a common limitation of public databases.[10,19] Lack of validation with an external data set currently prevents this model from being predictive. In future work, we aim to validate this statistical model using CT scans from other data sources and possibly develop a predictive model that could be useful in the clinic. In addition, we can include other important covariates such as KPS to assess their interaction with survival.

We only explored 0D topological feature curves in our analysis. We had found similar trends using other topological features curves in predicting survival. However, adding the moments of distribution of the other three feature curves would add up to 12 additional variables to consider, which we do not believe our study is adequately powered to assess. We chose to limit our focus to 0D features as they had the strongest association with survival.

6 | CONCLUSION

Most of the aforementioned issues can be resolved by extending our pipeline to study additional CT scans from NSCLC patients and additional clinical covariates across new databases. This would allow us to control for other clinical confounders and provide greater statistical power to assess moments of distribution of higher dimensional topological feature curves. Such an approach would allow for us validate our model and potentially develop a predictive one.

After initial therapy, patients with lung cancer obtain follow-up CT scans to survey tumor recurrence. Local and regional failure occur, respectively, when the primary tumor site and nearby lesions regrow after therapy. As many of the curative intent therapies include primary-tumor-directed therapy (radiation or surgery), it would be worthwhile in the future to also consider local control or local progression free survival in addition to overall survival. Long term, we envision predictive modeling that incorporates relevant clinical and radiomics variables to personalize anticancer therapy for each patient. The topological feature curve summary statistics may be appropriate variables to include in such models. As our variable only requires a CT scan and appropriate segmentation, this tool would provide universal benefit to all healthcare practice types.

ACKNOWLEDGMENTS

We would like to thank Theory Division for their constructive feedback and valuable input in this project. Funding for this project was provided by the Case Comprehensive Cancer Center medical student summer research training grant. J.G.S. was supported by the National Cancer Institute of the National Institutes of Health, under grant R37 CA244613.

An open-source, reproducible pipeline with all data and code are available at https://github.com/eashwarsoma/TDA-Lung-Phom-Reproducible with detailed instructions.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

REFERENCES

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577.
2. Liu Z, Wang S, Dong D, et al.. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;9(5):1303-1322.
3. AI B, Oncology CCC. *Comparative Oncology*. The Publishing House of the Romanian Academy; 2007.
4. Gordetsky J, Epstein J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn Pathol*. 2016;11.
5. Cámara PG. Topological methods for genomics: present and future directions. *Curr Opin Syst Biol*. 2017;1:95-101.
6. Lawson P, Sholl AB, Brown JQ, Fasy BT, Wenk C. Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Sci Rep*. 2019;9:1139.
7. Crawford L, Monod A, Chen AX, Mukherjee S, Rabadán R. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *J Am Stat Assoc*. 2019;115:1139-1150.
8. Oyama A, Hiraoka Y, Obayashi I, et al.. Hepatic tumor classification using texture and topology analysis of non-contrast-enhanced three-dimensional T1-weighted MR images with a radiomics approach. *Sci Rep*. 2019;9:8764.
9. Rizzo S, Botta F, Raimondi S, et al.. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp*. 2018;2.
10. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al.. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5.
11. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clinic Proceedings*. 2008;83:584-594.
12. Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. A roadmap for the computation of persistent homology. *EPJ Data Science*. 2017;6.
13. Zomorodian A. Proceedings of Symposia in Applied Mathematics. *Advances in Applied and Computational Topology*. 2012;70:1-39.
14. Ghrist R. Barcodes: The persistent topology of data. *Bull Am Math Soc*. 2007;45:61-76.
15. Gracia-Tabuenca Z, Díaz-Patiño JC, Arelio I, Alcauter S. Topological data analysis reveals robust alterations in the

whole-brain and frontal lobe functional connectomes in attention-deficit/hyperactivity disorder. *Eneuro*. 2020;7:ENEURO.0543-19.2020.
16. Clark K, Vendt B, Smith K, et al.. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. 2013;26:1045-1057.
17. Aerts HJWL, Wee L, Rios Velazquez E, et al. Data from NSCLC-Radiomics, 2019.
18. Bakr S, Gevaert O, Echegaray S, et al. Data for NSCLC Radiogenomics Collection, 2017.
19. Bakr S, Gevaert O, Echegaray S, et al.. A radiogenomic dataset of non-small cell lung cancer. *Sci Data*. 2018;5:180202.
20. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2012. ISBN 3-900051-07-0.
21. RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC; 2020.
22. Van Rossum G, Drake FL. *Python 3 Reference Manual*. Scotts Valley, CreateSpace; 2009.
23. Whitcher B, Schmid VJ, Thornton A. Working with the DICOM and NIfTI Data Standards in R. *J Stat Software*. 2011;44:1-28.
24. Whitcher B, Schmid VJ, Thornton A. Working with the DICOM and NIfTI data standards in R. *J Stat Software*. 2011;44:1-28.
25. Clayden J, Cox B, Jenkinson M. *RNifti: Fast R and C++ Access to NIfTI Images*. 2020. R package version 1.1.0.
26. Wickham H, François R, Henry L, Müller K. *dplyr: a grammar of data manipulation*. 2020. R package version 1.0.0.
27. Wickham H. The split-apply-combine strategy for data analysis. *J Stat Software*. 2011;40:1-29.
28. Wickham H. Reshaping data with the reshape package. *J Stat Software*. 2007;21.
29. Ushey K, Allaire J, Tang Y. *reticulate: Interface to 'Python'*. 2020. R package version 1.16.
30. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer; 2000.
31. Kassambara A. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. 2020. R package version 0.6.0.
32. Kassambara A, Kosinski M, Biecek P. *survminer: Drawing Survival Curves using 'ggplot2'*. 2020. R package version 0.4.7.
33. Yoshida K. *tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights*. 2020. R package version 0.11.2.
34. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag; 2016.
35. Wadhwa RR, Williamson DFK, Dhawan A, Scott JG. TDAstats: R pipeline for computing persistent homology in topological data analysis. *J Open Source Software*. 2018;3:860.
36. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020.
37. Urbanek S. *png: Read and write PNG images*. 2013. R package version 0.1-7.
38. Yu G. *ggplotify: Convert Plot to 'grob' or 'ggplot' Object*. 2020. R package version 0.0.5.
39. Kassambara A. *ggpubr: 'ggplot2' Based Publication Ready Plots*. 2020. R package version 0.4.0.
40. Iannone R, Cheng J, Schloerke B. *gt: Easily Create Presentation-Ready Display Tables*. 2020. R package version 0.2.2.
41. Hvitfeldt E. *paletteer: Comprehensive Collection of Color Palettes*. 2020. R package version 1.2.0.
42. Iannone R. *DiagrammeR: Graph/Network Visualization*. 2020. R package version 1.0.6.1.
43. Iannone R. *DiagrammeRsvg: Export DiagrammeR Graphviz Graphs as SVG*. 2016. R package version 0.1.
44. Horikoshi M, Tang Y. *ggfortify: Data Visualization Tools for Statistical Analysis Results*. 2018.
45. Ooms J. *rsvg: Render SVG Images into PDF, PNG, PostScript, or Bitmap Arrays*. 2020. R package version 2.1.
46. Auguie B. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. 2017. R package version 2.3.
47. Oliphant TE. *A guide to NumPy*. 1. Trelgol Publishing USA. 2006.
48. The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board. 3.3.0 ed. 2020.
49. Herz C, Fillion-Robin JC, Onken M, et al.. dcmqi: An Open Source Library for Standardized Communication of Quantitative Image Analysis Results Using DICOM. *Cancer Res*. 2017;77:e87-e90.
50. Yushkevich PA, Piven J, Hazlett HC, et al.. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*. 2006;31:1116-1128.
51. Mackin D, Fave X, Zhang L, et al.. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*. 2015;50:757-765.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Somasundaram E, Litzler A, Wadhwa R, Owen S, Scott J. Persistent homology of tumor CT scans is associated with survival in lung cancer. *Med Phys*. 2021;48:7043–7051.
https://doi.org/10.1002/mp.15255