medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Persistent Homology of Tumor CT Scans Predicts Survival In Lung Cancer

Eashwar Somasundaram[1,*], Adam Litzler[2], Raoul R. Wadhwa[3], and Jacob G. Scott[1,3,4,+]

[1]Case Western Reserve University School of Medicine, Cleveland, OH 44195, USA
[2]University of Colorado Boulder, Department of Mathematics, Boulder, CO 80309, USA
[3]Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA
[4]Taussig Cancer Institute, Departments of Translational Hematology and Oncology Research and Radiation Oncology, Cleveland, OH 44106, USA
[*]evs27@case.edu
[+]Correspondence: scottj10@ccf.org

ABSTRACT

Radiomics, the objective study of non-visual features in clinical imaging, has been useful in informing decisions in clinical oncology. However, radiomics currently lacks the ability to characterize the overall structure of the data. This field may benefit by incorporating persistent homology, a popular new algorithm that analyzes whole data structure. We hypothesized that persistent homology could be applied to lung tumor scans and predict clinical variables. We obtained computed tomography lung scans (n = 565) from the NSCLC-Radiomics and NSCLC-Radiogenomics datasets in The Cancer Imaging Archive. Segmentation data was used to create a cubical region centered on the primary tumor in each scan. For each scan, a cubical complex filtration based on Hounsfield units was generated. We created a feature curve that plotted the number of 0 dimensional topological features against each Hounsfield unit. The curve's first moment of the distribution was utilized as a summary statistic to predict survival in a Cox proportional hazards model. The first moment of the distribution is equivalent to the area under the curve of our topological feature curves (AUC). The Kruskal-Wallis H Test and a post-hoc Dunn's test with Bonferroni correction were used to test AUC differences among survival quartiles. After controlling for tumor image size, age, and stage, AUC, was associated with poorer survival (HR = 1.118; 95% CI = 1.026-1.218; $p = 0.01$). AUC was significantly higher for patients in the lowest survival quartile compared to the highest survival quartile ($p < 0.001$). We have shown that persistent homology can generate useful clinical correlates from tumor CT scans. Our 0-dimensional 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.

1 Introduction

The use of radiomics in tumor imaging has been an invaluable, noninvasive tool informing 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.

Since topological properties would focus on general tumor shape properties, they may be more robust to noise and 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]. While 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. Since we hypothesized earlier that an increased number of topological features may correlate

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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 0D topological feature curve, our way 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. area under the curve) is significantly associated with worse survival.

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[10]. Counting topological features on an image requires a binary black and white image, so we cannot count the topological features from the grayscale 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.

For each CT scan, we create a series of binary images (i.e. cubical complexes), one for each Hounsfield unit filtration value. For example, if the filtration value is -900 Hounsfield units, 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 Hounsfield unit filtration values across which the feature is found.

We show this filtration process with slice 13 of scan 1 in figure 1A. This slice is shown as it demonstrates a potential connection between tissue characteristics (calcifications) and topological features. At low Hounsfield unit 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 features. Two-dimensional features are not shown in figure 1 as they only appear when considering the cubical complexes of all the slices together as a three-dimensional structure. Intuitively, they can be thought of as pockets of white pixels. Calcifications would likely manifest as 1D or 2D features late in the filtration, which is shown when the Hounsfield unit threshold is 200 in figure 1A.

Topological barcodes are one of the most popular ways to visualize persistent homology[11]. Each topological feature in a barcode diagram is given a color to represent its dimension and a horizontal bar that spans the Hounsfield unit filtration values where the feature can be found. The barcode diagram in figure 1B represents the entire cubical complex persistent homology of scan 1. Since 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 normalized Hounsfield unit filtration value. 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 1C represent all feature curves in scan 1.

We use the moments of the distribution of the resulting 0D topological feature curve as predictor variables for our survival analysis. We used raw moments. The exact mathematical calculation is described in Table 3 in the Supplement.

Similar topological summary statistics have been described in analyzing brain connectomes of individuals with ADHD using 0D Vietoris-Rips complex persistent homology features[12]. Likewise, we also focus our analysis on 0D features; however, we generate our topological features using cubical complexes.

3 Results

Table 1 gives a descriptive overview of the two lung scan data sets. These data sets were obtained from The Cancer Imaging Archive[13]. The NSCLC-Radiomics set (n = 421) had more patients compared to the NSCLC Radiogenomics set (n = 138)[14,15].

The NSCLC Radiogenomics had a greater proportion of patients who died during its study. There was no significant difference in age between the Radiomics cohort (68.05 $\pm$ 10.09) and the Radiogenomics cohort (69.27 $\pm$ 8.79, t = 1.27, p = 0.206). Stage had a different distribution between the two cohorts (Stage I tumors 22.1% vs 63.0%, χ^2 = 135, p < .001). There were no significant differences in sex proportion between the two data sets (χ^2 = 1.38, p = 0.239). Males were most prevalent in both sets. Moment 1 of the 0D feature curve (Z = 28592, p = 0.782) and tumor image size (Z = 30544, p = 0.36) were not significantly different between the two data sets. Both of these variables were significant in predicting survival in the univariate Cox hazard model.

	Whole Cohort	NSCLC-Radiogenomics	NSCLC-Radiomics	p-value	Proportion Missing

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Total Sample Size	559	138	421		
Vital Status (% Dead)	424 (75.8)	51 (37.0)	373 (88.6)	<0.001	0.0
Moment 1	24.23 [5.57, 70.40]	21.07 [5.22, 69.19]	26.26 [5.64, 71.48]	0.782	0.0
Moment 2	9.49×10^3 [7.72×10^2, 5.99×10^4]	1.37×10^4 [8.69×10^2, 1.64×10^5]	8.74×10^3 [7.06×10^2, 4.55×10^4]	0.032	0.0
Moment 3	3.92×10^6 [1.31×10^5, 6.56×10^7]	9.12×10^6 [1.63×10^5, 4.21×10^8]	3.29×10^6 [9.85×10^4, 4.16×10^7]	0.003	0.0
Moment 4	1.97×10^9 [2.25×10^7, 8.76×10^{10}]	6.98×10^9 [4.15×10^7, 1.21×10^{12}]	1.38×10^9 [1.36×10^7, 4.62×10^{10}]	0.001	0.0
Tumor Image Size	4.60×10^4 [1.19×10^4, 1.30×10^5]	4.51×10^4 [1.22×10^4, 1.63×10^5]	4.60×10^4 [1.19×10^4, 1.24×10^5]	0.364	0.0
Age	68.36 (9.78)	69.27 (8.79)	68.05 (10.09)	0.206	3.9
Stage				<0.001	0.2
Stage I	180 (32.3)	87 (63.0)	93 (22.1)		
Stage II	66 (11.8)	26 (18.8)	40 (9.5)		
Stage IIIa	131 (23.5)	20 (14.5)	111 (26.4)		
Stage IIIb	177 (31.7)	1 (0.7)	176 (41.9)		
Stage IV	4 (0.7)	4 (2.9)	0 (0.0)		
Sex (% Female)	166 (29.7)	35 (25.4)	131 (31.1)	0.239	0.0

Table 1. Descriptive statistics of the two cohorts of patient scans used in this study. All patient scans were downloaded from The Cancer Imaging Archive[13]. The Radiomics set had 421 segmentable tumor scans, and the Radiogenomics set had 144 total segmentable scans. 6 patients were excluded from the Radiogenomics set since they had stage 0 cancer. No stage 0 patient died, so the stage 0 variable did not converge in the Cox Hazard models (Table 5 in the supplement). The moments of the 0D feature curves are presented as median [IQR] since the data was not normally distributed. Tumor image size was also not normally distributed and also presented as median [IQR]. All other data are presented as mean (SD) or as a proportion. Statistical comparisons between the two cohorts were performed using Wilcoxon rank-sum test for the non normally distributed data. Student's t-test was used for the other continuous data. Chi-squared test was used for the categorical stage and sex data.

3.1 Discrete Analysis of 0D Topological Feature Curves

We first performed a discrete analysis comparing survival groups by their moments of distribution of the 0D feature curves. Patients were placed in survival groups based on their survival quartile in the combined study cohort. This discrete analysis was performed to explore possible visual differences in how our new 0D topological feature curve metric interacted with survival. Figure 6 in the supplement shows that poorer survival quartiles had taller 0D feature curves. Figure 7 in the supplement quantifies the difference in these curves showing that poorer survival quartiles had larger moments of the distribution. Fully detailed statistical comparisons among the moments of distribution values are found in table 4 in the supplement.

3.2 Cox Proportional Hazard Analysis of 0D Topological Feature Curves

The discrete analysis was performed primarily for intuition and data visualization. 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 figure 9 in the supplement. Table 2 shows the results of our Cox proportional hazard model.

Cox Proportional Hazard Model

	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 vs Female	1.229 (0.994-1.52)	0.0570	1.128 (0.898-1.417)	0.3000
Scaled Moment 1	1.019 (1.005-1.033)	0.0082	1.118 (1.026-1.218)	0.0110
Scaled Moment 2	1.004 (0.982-1.026)	0.7400	0.766 (0.53-1.106)	0.1600
Scaled Moment 3	0.992 (0.94-1.047)	0.7700	1.282 (0.412-3.991)	0.6700
Scaled Moment 4	0.985 (0.893-1.088)	0.7700	0.995 (0.369-2.681)	0.9900

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Scaled Tumor Image Size	1.015 (1.003-1.026)	0.0100	1.014 (0.983-1.046)	0.3800
Stage II vs I	1.249 (0.884-1.765)	0.2100	1.028 (0.71-1.489)	0.8800
Stage IIIa vs I	1.695 (1.304-2.203)	8.0×10^{-5}	1.742 (1.317-2.305)	0.0001
Stage IIIb vs I	1.626 (1.272-2.077)	0.0001	1.483 (1.127-1.951)	0.0049
Stage IV vs I	1.279 (0.406-4.028)	0.6700	1.825 (0.572-5.818)	0.3100

Table 2. Both univariate and multivariate Cox hazard models show moment 1 of the 0D feature curve is associated with poorer survival. Data is 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 vs stage I variable did not converge in the model (table 5 in supplement).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

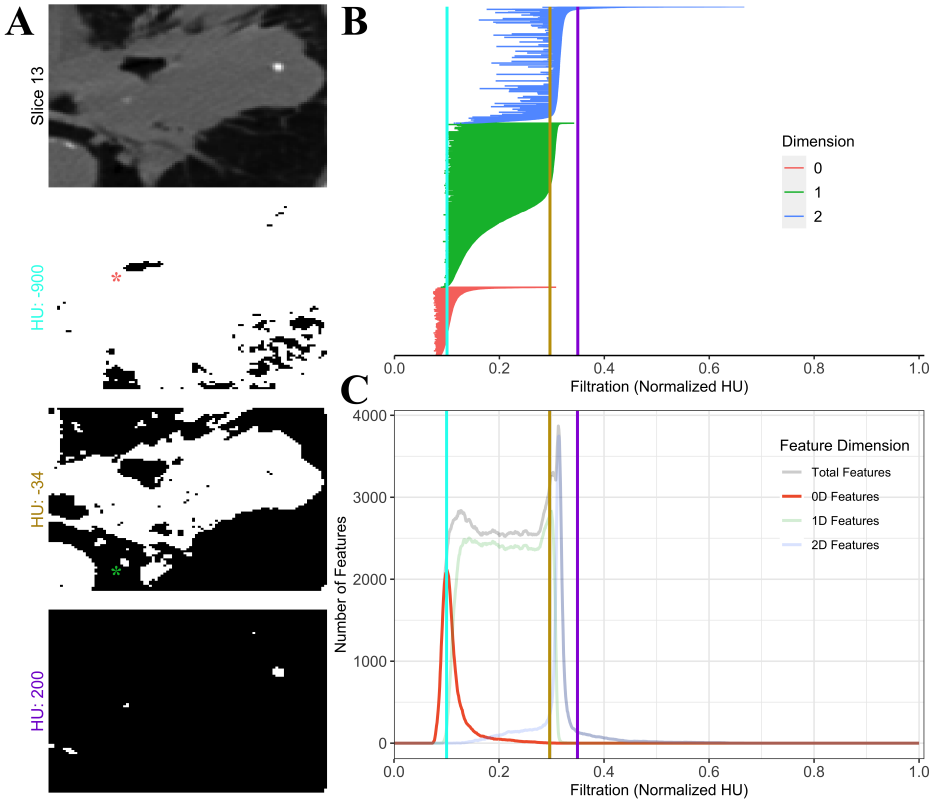

Figure 1. Methodology to generate cubical complexes. Slice 13 of CT scan 1 from the Radiomics dataset is shown in panel A. Each pixel value in a CT scan is described by a Hounsfield unit, which typically range from -1024 to 3071. We create a cubical complex by selecting a Hounsfield value as a filter. Any pixel below or equal to this filter is colored black. Any pixel above is colored white. This binary image is considered a 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 three-dimensional structure. Intuitively, they can be thought of as pockets of white pixels. When the filtration reaches the maximum value, the whole image is colored black ending the filtration process. A barcode diagram (panel B) represents persistent homology by giving each topological feature a barcode. Color represents dimension, and the range represents the Hounsfield filtration unit range during which the feature existed. The Hounsfield unit horizontal axis in the barcode diagram has been normalized zero to one. The colored vertical lines represent the normalized Hounsfield units from the filtration values shown in the binary images. We create three topological feature curves (panel C) 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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, which also represents the AUC, 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 2 shows a forest plot of the Cox model. The nonsignificant Stage IV vs Stage I predictor variable was removed as its wide confidence interval interfered with data visualization.

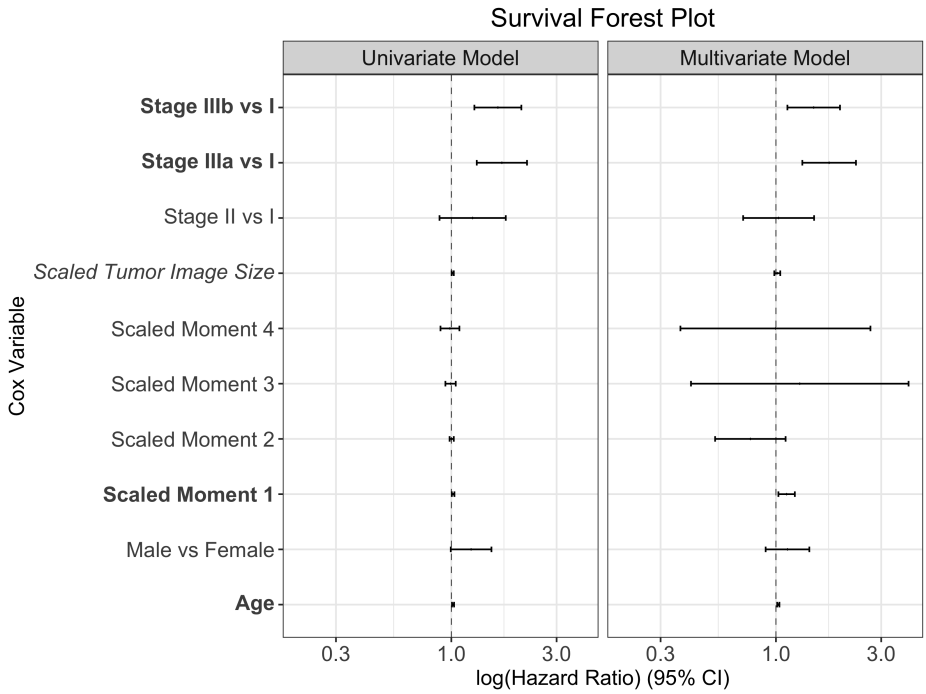

Figure 2. Survival forest plot shows moment 1 predicts poorer survival even after controlling for known tumor prognosticators such as stage. The horizontal axis is log transformed for better data visualization. Span of each data point represents the 95% CI. The Stage IV vs Stage I data point had a very wide CI, so it was removed from this plot for data visualization. The Stage IV vs 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.

We used Kaplan-Meier curves (figure 3) 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 ($\chi^2 = 12.5$; $p = 0.0024$).

4 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 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 since 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

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

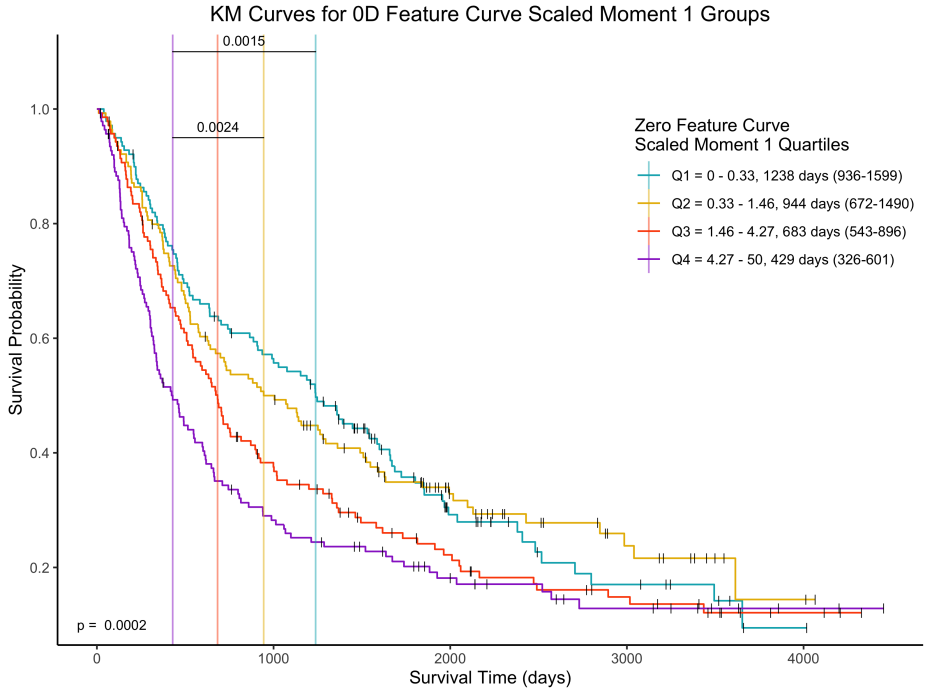

Figure 3. Patients with larger moment 1 values of the 0D feature curve had poorer survival. Patients were divided into quartiles based on the moment 1 value of the 0D topological feature curve. Q1 through Q4 represent Quartile 1 through Quartile 4. The exact quartile values and median survival of each quartile is 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

30% of the variance in moment 1 of the 0D topological feature curve was not explainable by tumor image size alone (figure 10 in the supplement).

Despite these promising results, there remain limitations in our study. Though the pipeline worked well across two independent batches, 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[16]. While 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. In addition, while we know the analyzed scans were for planning radiotherapy or surgery, all of the patients' prior clinical history is a black box, which is a common limitation of public databases[17,18].

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.

Most of these issues can be resolved by extending our pipeline to study additional CT scans from NSCLC patients with known clinical outcome across different 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.

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. Since our variable only requires a CT scan and appropriate segmentation, this tool would provide universal benefit to all healthcare practice types.

5 Methods

A publicly available pipeline using R (v3.6.1) and Python (v3.7.6) code was built for this project[19–21]. All images and segmentation objects were obtained from The Cancer Imaging Archive[13]. Within TCIA, the NSCLC-Radiomics (n = 422) and NSCLC Radiogenomics (n = 211) cohorts provided CT scans and clinical data[14,15,17,18]. Of the 633 total scans, 565 also had segmentation data, which was necessary for our pipeline. Figure 4 shows the exclusion and inclusion criteria for each analysis.

Currently, R has a limitation in reading .dcm segmentation files. To work around this issue, we first converted the .dcm segmentation files into the NIfTI file format using the dcmqi open source Bash library[22].

In R, we used `oro.dicom` v0.5.3, `oro.nifti` v0.10.3, and `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 ITK-SNAP v3.8.0 as the DICOM viewer to visually ensure our cubical segmentation properly highlighted the tumor region of interest[49].

Figure 5 shows an overview of the data pipeline. Within R, a cubical region that encapsulated the tumor was delineated using the NIfTI segmentation files. This data object was exported to Python to compute cubical complex persistence homology (functionality in R not present at time of analysis). Cubical complex persistence homology was computed in Python and exported as .csv files to be processed in R. Within R, the raw persistence homology was transformed into topological feature curves, which plot the topological features against each Hounsfield filtration value. The first four moments of distribution were calculated for specifically 0D topological feature curves. These moments were used as predictor variables for our survival analysis.

Unpaired t-test statistics (for normally distributed data) and Wilcoxon rank sum tests (for non-normally 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. In the discrete analysis, all moments of distribution had a skewed exponential distribution, so non-parametric statistical tests were performed. Kruskall-Wallis H test was used to compare the median

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

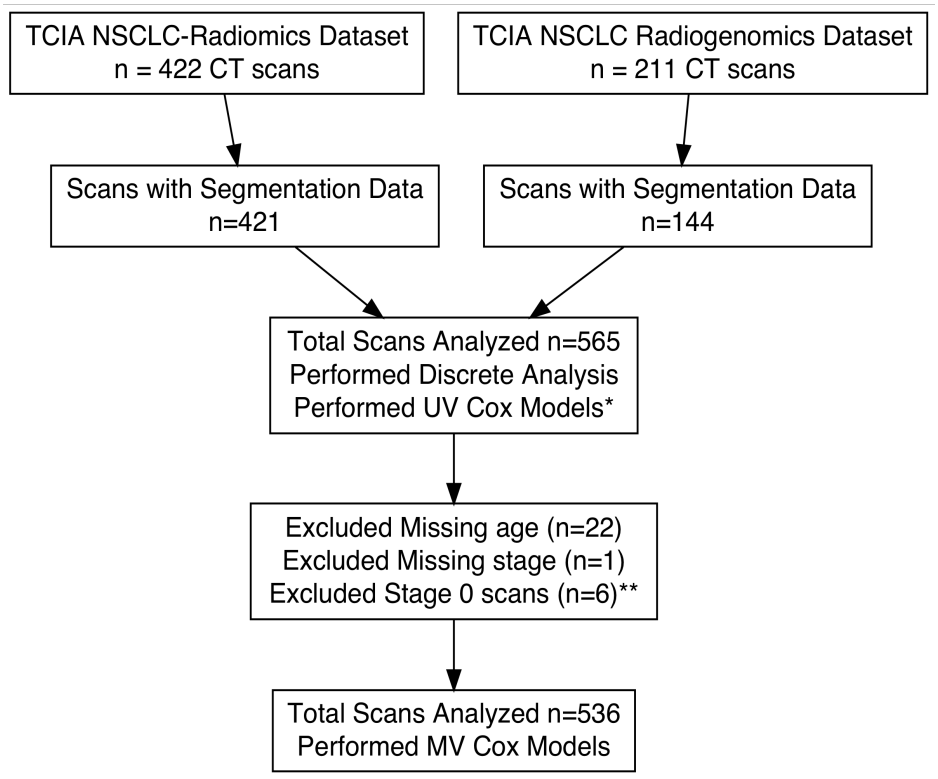

Figure 4. Flowchart of tumor scan inclusion and exclusion. 565 tumors had both segmentation data and CT scans allowing for computation of cubical complexes in our pipeline. *The univariate Cox model used all 565 tumor data assuming no NAs were present. However, 22 patients were missing age information, and 1 patient was missing stage information. These patients were excluded in their respective univariate Cox analysis. **The multivariate Cox model excluded 6 additional patients who were stage 0 since the Cox models did not converge (table 5 in the supplement).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

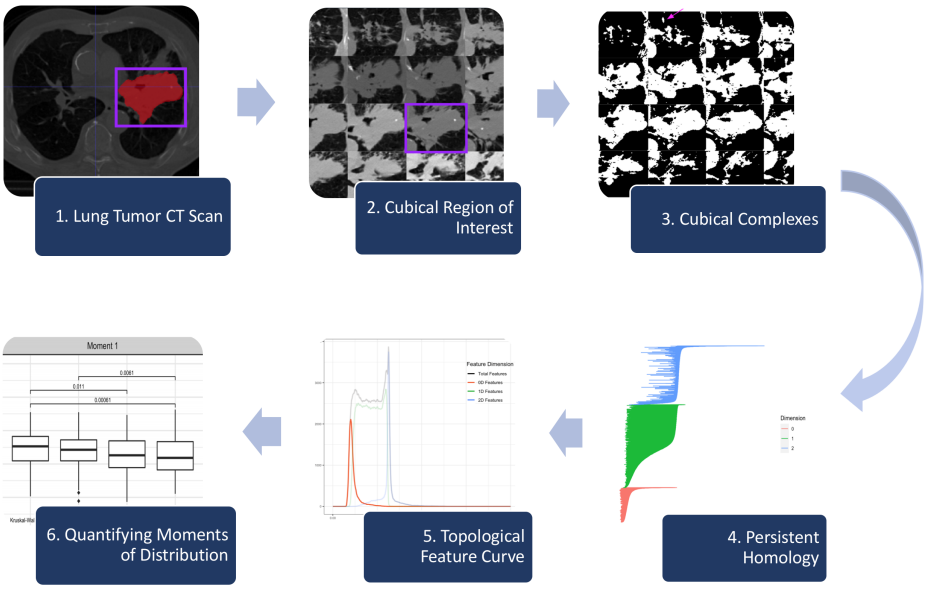

Figure 5. Overview of data pipeline. Using the tumor segmentation data, a cubical region that encapsulated the tumor was isolated from each scan. This data object was created in R and passed on to Python to compute persistent homology. Persistent homology was computed using cubical complexes. The pink arrow in the third panel points to an example of a 1D feature. The full persistent homology can be described in a bar code diagram where the horizontal axis represents the Hounsfield filtration value. Each bar in the diagram represents a single topological feature, and its range represents persistence. We transform this object into topological feature curves, which plot topological features against the normalized Hounsfield unit filtration values. Finally, we calculated the moments of distributions for 0D feature curves to be predictor variables in our survival analysis.

moments values of all survival groups, and a *post hoc* Dunn's test with Bonferroni correction was used to make pairwise comparisons.

For the continuous analysis, univariate and multivariate Cox proportional hazard modeling were performed for the moments, tumor image size, cancer stage, age, and sex. We included tumor image size as a covariate in the multivariable Cox Hazard model to control for potential confounding with our 0D topological feature curve. Increasing the number pixels of an image may increase the topological features without necessarily changing the overall topological structure of the tumor image. Tumor image size was calculated as the product of number of pixels per slice and number of slices. 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 since none of the 6 patients with stage 0 cancer died during the study, so these patients were excluded (table 5 in the supplement).

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. All post hoc statistical tests p-values received Bonferroni correction. Summary statistics are stated as mean $\pm$ SD or median [IQR]. All code to reproduce results and figure is publicly available on `https://github.com/eashwarsoma/TDA-Lung-Phom-Reproducible` with detailed instructions.

References

1. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **278**, 563–577, DOI: 10.1148/radiol.2015151169 (2016).

2. Liu, Z. *et al.* The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. *Theranostics* **9**, 1303–1322 (2019).

3. AI, B. & Oncology, C. C. C. *Comparative Oncology* (Bucharest: The Publishing House of the Romanian Academy, 2007).

4. Gordetsky, J. & Epstein, J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn. Pathol.* **11**, DOI: 10.1186/s13000-016-0478-2 (2016).

5. Cámara, P. G. Topological methods for genomics: present and future directions. *Curr Opin Syst Biol* **1**, 95–101 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

6. Lawson, P., Sholl, A. B., Brown, J. Q., Fasy, B. T. & Wenk, C. Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology. *Sci Rep* **9**, 1139 (2019).

7. Crawford, L., Monod, A., Chen, A. X., Mukherjee, S. & Rabadán, R. Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *J. Am. Stat. Assoc.* 1–12, DOI: 10.1080/01621459.2019.1671198 (2019).

8. Oyama, A. *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. Reports* **9**, DOI: 10.1038/s41598-019-45283-z (2019).

9. Rizzo, S. *et al.* Radiomics: the facts and the challenges of image analysis. *Eur. Radiol. Exp.* **2**, DOI: 10.1186/s41747-018-0068-z (2018).

10. Otter, N., Porter, M. A., Tillmann, U., Grindrod, P. & Harrington, H. A. A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**, DOI: 10.1140/epjds/s13688-017-0109-5 (2017).

11. Ghrist, R. Barcodes: The persistent topology of data. *Bull. Am. Math. Soc.* **45**, 61–76, DOI: 10.1090/s0273-0979-07-01191-3 (2007).

12. Gracia-Tabuenca, Z., Díaz-Patiño, J. C., 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* **7**, ENEURO.0543–19.2020, DOI: 10.1523/eneuro.0543-19.2020 (2020).

13. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit. Imaging* **26**, 1045–1057 (2013).

14. Aerts, H. J. W. L. *et al.* Data from nsclc-radiomics, DOI: 10.7937/K9/TCIA.2015.PF0M9REI (2019).

15. Bakr, S. *et al.* Data for nsclc radiogenomics collection, DOI: 10.7937/K9/TCIA.2017.7HS46ERV (2017).

16. Mackin, D. *et al.* Measuring computed tomography scanner variability of radiomics features. *Investig. Radiol.* **50**, 757–765, DOI: 10.1097/rli.0000000000000180 (2015).

17. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **5**, DOI: 10.1038/sdata.2018.202 (2018).

18. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, DOI: 10.1038/ncomms5006 (2014).

19. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2012). ISBN 3-900051-07-0.

20. RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA (2020).

21. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

22. Herz, C. *et al.* dcmqi: An open source library for standardized communication of quantitative image analysis results using DICOM. *Cancer Res.* **77**, e87–e90, DOI: 10.1158/0008-5472.can-17-0336 (2017).

23. Whitcher, B., Schmid, V. J. & Thornton, A. Working with the DICOM and NIfTI data standards in R. *J. Stat. Softw.* **44**, 1–28 (2011).

24. Whitcher, B., Schmid, V. J. & Thornton, A. Working with the DICOM and NIfTI data standards in R. *J. Stat. Softw.* **44**, 1–28 (2011).

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. Softw.* **40**, 1–29 (2011).

28. Wickham, H. Reshaping data with the reshape package. *J. Stat. Softw.* **21** (2007).

29. Ushey, K., Allaire, J. & Tang, Y. *reticulate: Interface to 'Python'* (2020). R package version 1.16.

30. Terry M. Therneau & Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model* (Springer, New York, 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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 New York, 2016).

35. Wadhwa, R. R., Williamson, D. F. K., Dhawan, A. & Scott, J. G. TDAstats: R pipeline for computing persistent homology in topological data analysis. *J. Open Source Softw.* **3**, 860, DOI: 10.21105/joss.00860 (2018).

36. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (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, T. E. *A guide to NumPy*, vol. 1 (Trelgol Publishing USA, 2006).

48. The GUDHI Project. *GUDHI User and Reference Manual* (GUDHI Editorial Board, 2020), 3.3.0 edn.

49. Yushkevich, P. A. *et al.* User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage* **31**, 1116–1128, DOI: 10.1016/j.neuroimage.2006.01.015 (2006).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Appendices

A Supplemental Tables and Figures

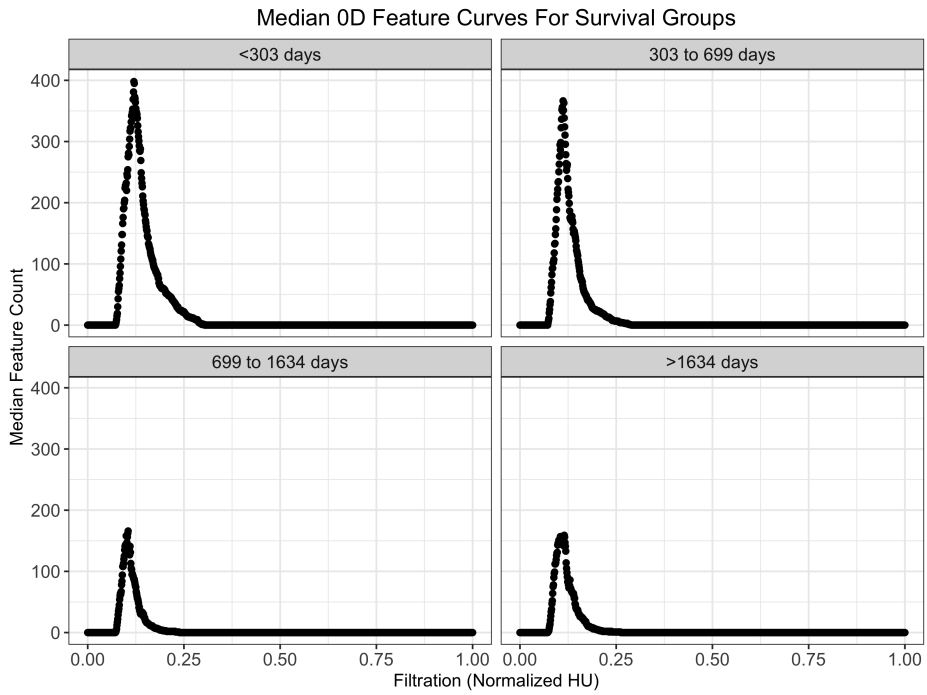

Figure 6. Median 0D feature curves are larger for patients in poorer survival quartiles. Each survival group had approximately one quarter of the combined study cohorts (n = 145, n = 138, n = 141, n = 141, from worst to best survival). Within each survival group, the median 0D feature count value for each Hounsfield units was collected into one vector and plotted. The Hounsfield units were normalized to 0 to 1.

Moments of Distribution Formulas

Moment	Formula
Moment 1	$E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i$
Moment 2	$E[X^2] = \frac{1}{N}\sum_{i=1}^{N} X_i^2$
Moment 3	$E[X^3] = \frac{1}{N}\sum_{i=1}^{N} X_i^3$
Moment 4	$E[X^4] = \frac{1}{N}\sum_{i=1}^{N} X_i^4$

Table 3. We used the raw moments of distribution to analyze our 0D feature curves. X refers to the vector of values represented by the vertical axis in our 0D topological feature curves. N refers to the number of values in that vector. E refers to the expected value of a distribution.

Comparing 0D Feature Curve Moments by Survival Group

Group 1	Group 2	Dunn's z-statistic	adjusted p-value
Moment 1, KW H-statistic = 28.25, p = 0.0000032			
<303 days	303 to 699 days	-1.72	0.520000
<303 days	699 to 1634 days	-4.33	0.000089
<303 days	>1634 days	-4.51	0.000040
303 to 699 days	699 to 1634 days	-2.57	0.061000
303 to 699 days	>1634 days	-2.75	0.036000

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

699 to 1634 days	>1634 days	-0.17	1.000000
Moment 2, KW H-statistic = 21.03, p = 0.0001			
<303 days	303 to 699 days	-1.28	1.000000
<303 days	699 to 1634 days	-3.47	0.003100
<303 days	>1634 days	-4.01	0.000370
303 to 699 days	699 to 1634 days	-2.15	0.190000
303 to 699 days	>1634 days	-2.68	0.044000
699 to 1634 days	>1634 days	-0.53	1.000000
Moment 3, KW H-statistic = 18.49, p = 0.00035			
<303 days	303 to 699 days	-1.05	1.000000
<303 days	699 to 1634 days	-3.13	0.011000
<303 days	>1634 days	-3.77	0.000990
303 to 699 days	699 to 1634 days	-2.04	0.250000
303 to 699 days	>1634 days	-2.68	0.045000
699 to 1634 days	>1634 days	-0.64	1.000000
Moment 4, KW H-statistic = 17.64, p = 0.00052			
<303 days	303 to 699 days	-0.93	1.000000
<303 days	699 to 1634 days	-3.00	0.016000
<303 days	>1634 days	-3.66	0.001500
303 to 699 days	699 to 1634 days	-2.04	0.250000
303 to 699 days	>1634 days	-2.69	0.043000
699 to 1634 days	>1634 days	-0.66	1.000000

Table 4. Statistical comparisons between each survival group for each moment. Kruskal-Wallis H test was used to compare survival groups within each moment variable. Post hoc Dunn's test was performed to make pairwise comparison between survival groups. Bonferroni adjusted p-values are shown.

Supplemental Cox Proportional Hazard Model

	Univariate Model HR	p-value	Multivariate Model HR	p-value
Age	1.018 (1.007-1.029)	0.0018	1.025 (1.013-1.037)	4.0×10^{-5}
Male vs Female	1.212 (0.98-1.499)	0.0770	1.128 (0.898-1.417)	0.3000
Scaled Moment 1	1.019 (1.006-1.033)	0.0054	1.118 (1.026-1.218)	0.0110
Scaled Moment 2	1.005 (0.983-1.027)	0.6900	0.766 (0.53-1.106)	0.1600
Scaled Moment 3	0.994 (0.941-1.049)	0.8100	1.282 (0.412-3.991)	0.6700
Scaled Moment 4	0.988 (0.895-1.09)	0.8100	0.995 (0.369-2.681)	0.9900
Scaled Tumor Image Size	1.015 (1.003-1.026)	0.0100	1.014 (0.983-1.046)	0.3800
Stage 0 vs I	0 (0-Inf)	0.9900	0 (0-Inf)	0.9900
Stage II vs I	1.249 (0.884-1.765)	0.2100	1.028 (0.71-1.489)	0.8800
Stage IIIa vs I	1.695 (1.304-2.203)	8.0×10^{-5}	1.742 (1.317-2.305)	0.0001
Stage IIIb vs I	1.626 (1.272-2.077)	0.0001	1.483 (1.127-1.951)	0.0049
Stage IV vs I	1.279 (0.406-4.028)	0.6700	1.825 (0.572-5.818)	0.3100

Table 5. Cox table with stage 0 patients not excluded. Stage 0 patients were excluded in the original analysis as the stage 0 vs stage I variable did not converge since no stage 0 patients died. Including stage 0 patients does not impact any of the conclusions of the original analysis.

Acknowledgements

We would like to thank Steph Owen and the rest of 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

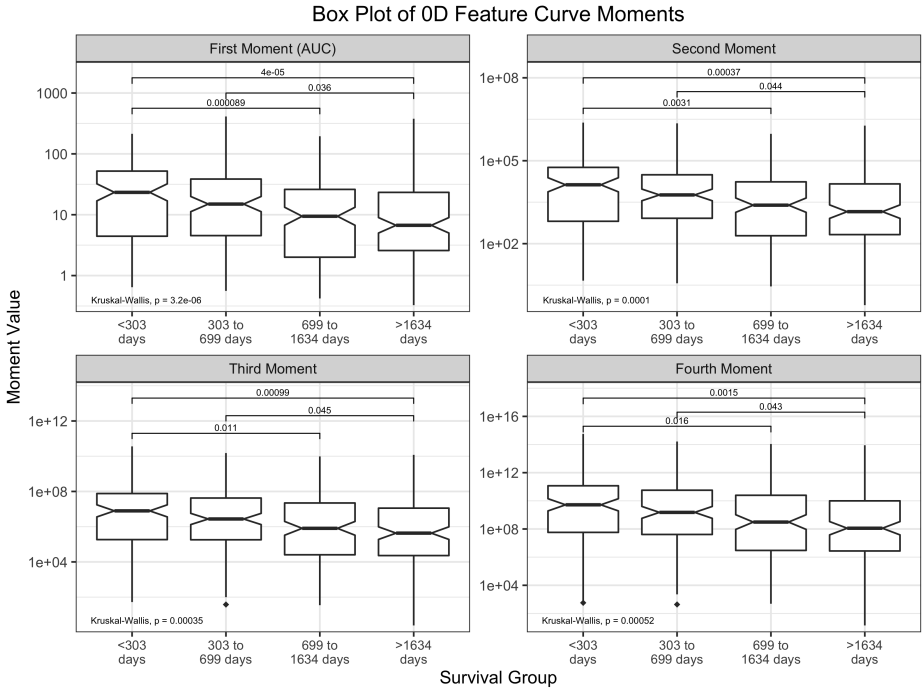

Figure 7. 0D feature curve moments are significantly higher for poorer survival quartiles as shown by box plots. The moments of distribution were analyzed for each survival group's 0D feature curves. For data visualization, the log of the moment values are presented in the vertical axis. Each facet's vertical axis has a different scale. Points on the plot represent outliers. Notches in the box represent a 95% confidence interval around the median. Kruskal-Wallis H Test was performed to compare median value for each moment among the survival groups, and post hoc Dunn's test with Bonferroni correction was performed to compare pairs of survival groups. Exact median moment values and interquartile ranges are provided in the supplemental figure 8. Exact Kruskal-Wallis and post hoc Dunn's test comparisons are found in table 4 in the supplement.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Survival Group Moment Distributions				
	<303 days	303 to 699 days	699 to 1634 days	>1634 days
Moment 1				
0%	0.644	0.56	0.421	0.326
25%	4.430	4.54	2.000	2.580
50%	23.300	15.00	9.430	6.700
75%	52.100	38.60	26.100	23.300
100%	214.000	411.00	195.000	378.000
Moment 2				
0%	4.590	3.69	2.840	0.596
25%	645.000	830.00	194.000	213.000
50%	1.4×10^4	5.8×10^3	2.5×10^3	1.4×10^3
75%	5.8×10^4	3.1×10^4	1.7×10^4	1.4×10^4
100%	2.4×10^6	2.2×10^6	9.4×10^5	1.9×10^6
Moment 3				
0%	52.700	38.00	34.900	2.470
25%	1.8×10^5	1.8×10^5	2.5×10^4	2.3×10^4
50%	7.9×10^6	2.7×10^6	8.1×10^5	4.3×10^5
75%	7.6×10^7	4.2×10^7	2.2×10^7	1.1×10^7
100%	3.6×10^{10}	1.5×10^{10}	9.7×10^9	1.2×10^{10}
Moment 4				
0%	569.000	434.00	488.000	14.000
25%	5.7×10^7	4.1×10^7	3.0×10^6	2.7×10^6
50%	5.2×10^9	1.5×10^9	3.1×10^8	1.1×10^8
75%	1.2×10^{11}	5.7×10^{10}	2.5×10^{10}	1.0×10^{10}
100%	5.7×10^{14}	1.6×10^{14}	1.1×10^{14}	8.9×10^{13}

Figure 8. Groups were divided into quartiles based on survival. Each 0D feature curve was analyzed for the first four moments of distribution. The range of the moment values for each group is provided here. Color depth represents magnitude of moment value.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

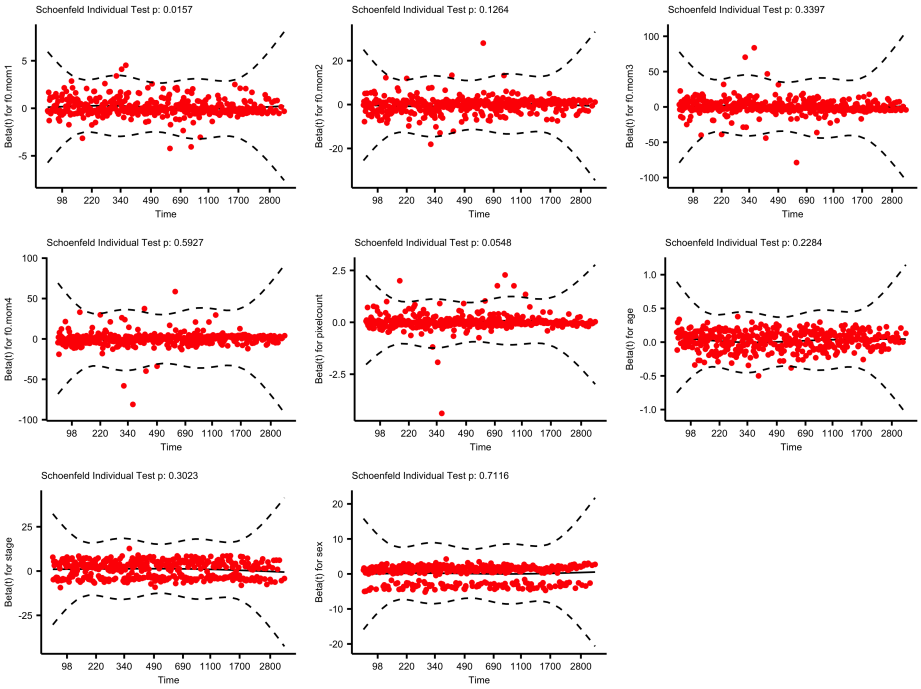

Figure 9. Schoenfeld residuals were visually appraised to ensure assumptions of proportionality were met. While moment 1's Schoenfeld residuals in the figure are significant, we believe it is a chance result as visually there does not seem to be a trend in the residuals over time. The global Schoenfeld residuals were also nonsignificant

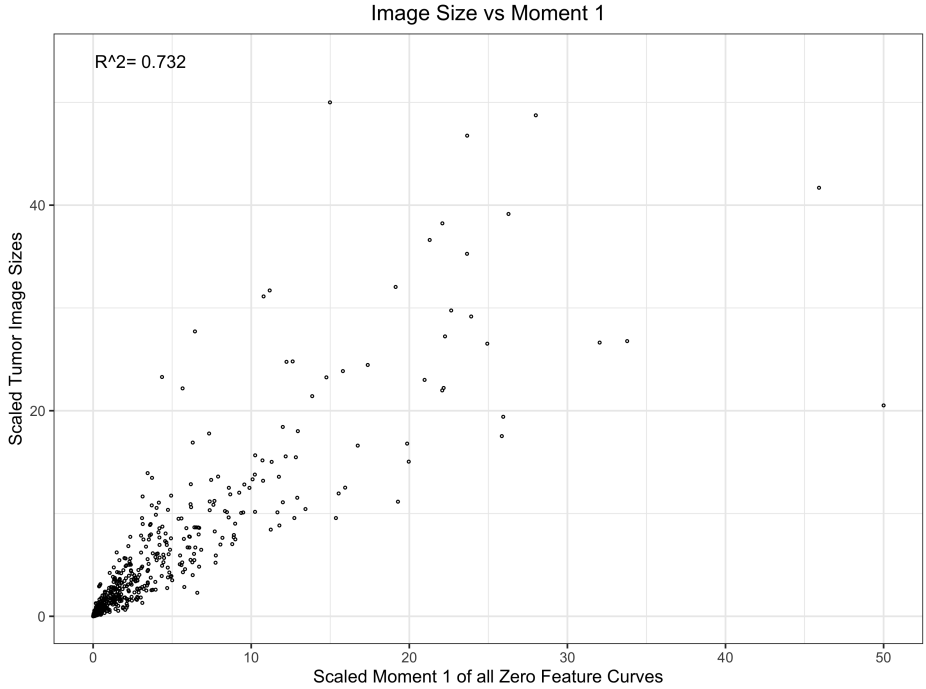

Figure 10. Correlation plot comparing tumor image size and moment 1 of 0D topological feature curves. Increasing the image size is expected to increase the topological features. Tumor image size was calculated by multiplying the number of pixels per CT scan slice by the number of slices. Both variables were linearly scaled to 0 to 50. We generated a correlation plot to check how much variance provided by moment 1 of the 0D feature curve was unexplained by tumor image size. About 30% of the variance in moment 1 is unexplained by tumor image size. Despite the large overlap, moment 1 remained significant even after controlling for tumor image size in our Cox model.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.06.20244863; this version posted December 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Author contributions statement

E.S. conceived the project. E.S., A.L., and R.W. designed project plan. E.S. and A.L. wrote code for the project. E.S. wrote manuscript. J.S. supervised all previous steps. All authors reviewed the manuscript.