Title

Deep segmentation networks predict survival of non-small cell lung cancer.

Journal

Scientific Reports, 9(1)

Authors

Baek, Stephen
Baek, Stephen
He, Yusen
et al.

Publication Date

2019-11-21

DOI

10.1038/s41598-019-53461-2

SCIENTIFIC REPORTS
nature research

OPEN

Deep segmentation networks predict survival of non-small cell lung cancer

Stephen Baek[1,2,3,7], Yusen He[1,7], Bryan G. Allen[2], John M. Buatti[2], Brian J. Smith[5], Ling Tong[4], Zhiyu Sun[1], Jia Wu[6], Maximilian Diehn[6], Billy W. Loo[6], Kristin A. Plichta[2], Steven N. Seyedin[2], Maggie Gannon[2], Katherine R. Cabel[2], Yusung Kim[2*] & Xiaodong Wu[2,3*]

Non-small-cell lung cancer (NSCLC) represents approximately 80–85% of lung cancer diagnoses and is the leading cause of cancer-related death worldwide. Recent studies indicate that image-based radiomics features from positron emission tomography/computed tomography (PET/CT) images have predictive power for NSCLC outcomes. To this end, easily calculated functional features such as the maximum and the mean of standard uptake value (SUV) and total lesion glycolysis (TLG) are most commonly used for NSCLC prognostication, but their prognostic value remains controversial. Meanwhile, convolutional neural networks (CNN) are rapidly emerging as a new method for cancer image analysis, with significantly enhanced predictive power compared to hand-crafted radiomics features. Here we show that CNNs trained to perform the tumor segmentation task, with no other information than physician contours, identify a rich set of survival-related image features with remarkable prognostic value. In a retrospective study on pre-treatment PET-CT images of 96 NSCLC patients before stereotactic-body radiotherapy (SBRT), we found that the CNN segmentation algorithm (U-Net) trained for tumor segmentation in PET and CT images, contained features having strong correlation with 2- and 5-year overall and disease-specific survivals. The U-Net algorithm has not seen any other clinical information (e.g. survival, age, smoking history, etc.) than the images and the corresponding tumor contours provided by physicians. In addition, we observed the same trend by validating the U-Net features against an extramural data set provided by Stanford Cancer Institute. Furthermore, through visualization of the U-Net, we also found convincing evidence that the regions of metastasis and recurrence appear to match with the regions where the U-Net features identified patterns that predicted higher likelihoods of death. We anticipate our findings will be a starting point for more sophisticated non-intrusive patient specific cancer prognosis determination. For example, the deep learned PET/CT features can not only predict survival but also visualize high-risk regions within or adjacent to the primary tumor and hence potentially impact therapeutic outcomes by optimal selection of therapeutic strategy or first-line therapy adjustment.

According to the World Health Organization (WHO), lung cancer remains the leading cause of cancer-related deaths worldwide, with 2.1 million new cases diagnosed and 1.8 million deaths in 2018[1]. NSCLC accounts for 80–85% of lung cancer diagnoses[2] and the five-year survival rate of NSCLC remains low (23%) compared to other leading cancer sites such as colorectal (64.5%), breast (89.6%), and prostate (98.2%)[3]. Historically, the tumor, nodes, and metastases (TNM) staging system has served as the major prognostic factor in predicting therapeutic outcomes, but it does not differentiate responders and non-responders within the same stage[4]. The maximum and the mean of standard uptake values (SUV_{MAX} and SUV_{MEAN}) have been reported for their correlation with survival[5–7], but are of limited clinical value due to their unsatisfactory predictive power and lack of robustness[8,9].

[1]University of Iowa, Department of Industrial and Systems Engineering, Iowa City, IA, 52242, United States. [2]University of Iowa, Department of Radiation Oncology, Iowa City, IA, 52242, United States. [3]University of Iowa, Department of Electrical and Computer Engineering, Iowa City, IA, 52242, United States. [4]University of Iowa, Department of Business Analytics, Iowa City, IA, 52242, United States. [5]University of Iowa, Department of Biostatistics, Iowa City, IA, 52242, United States. [6]Stanford University, Stanford Cancer Institute, Palo Alto, CA, 94304, United States. [7]These authors contributed equally: Stephen Baek and Yusen He. *email: yusung-kim@uiowa.edu; xiaodong-wu@uiowa.edu

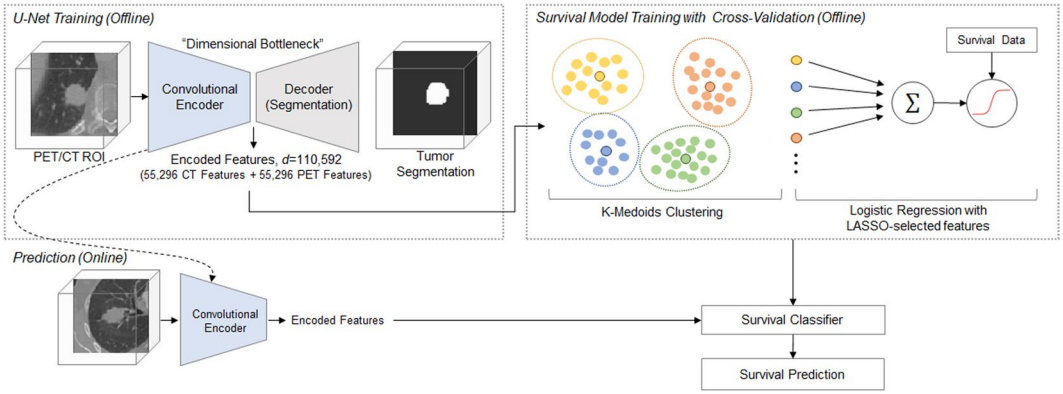

Figure 1. Schematic diagram of the survival prediction framework. The proposed framework consists of two major components: the U-Net segmentation network and the survival prediction model. The U-Net is trained with PET/CT images and corresponding physician contours but without survival-related information. The "dimensional bottleneck" at the middle of the U-Net produces latent variables summarizing image features (55,296 features from CT + 55,296 features from PET), which we hypothesize to be potentially relevant to cancer survival. These features are then clustered by *k*-medoids clustering approach in an unsupervised manner. Next, the LASSO method is used to select medoid features from the clusters based on their associations with survival. Last, a logistic regression model is trained for survival prediction, and survival prediction is performed when a new patient arrives with features extracted from the same U-Net.

Other prognostic markers have also been studied, including TLG, which incorporates metabolic tumor volume (MTV) and metabolic activity (TLG = MTV × SUV_{MEAN}). Reports[10–12] suggest that TLG may have better prognostic power than SUV_{MAX} or SUV_{MEAN}. These metrics, however, are not optimal and do not provide a comprehensive image-based analysis of tumors[13]. More recently, radiomics approaches, which employ semi-automated analysis based on a few hand-crafted imaging features describing intratumoral heterogeneity, demonstrated higher prognostic power[14,15]. However, these features still have limited predictive power ranging between 0.5 and 0.79 in terms of the area under the ROC (AUC)[15–17]. Recent literature on CNNs demonstrates their strong potential for cancer prognostication[18,19], however the clinical implications of deep learning remain questioned due to the limited interpretability of CNNs.

Here, we propose an interpretable and highly accurate framework to solve this problem by capitalizing on the unprecedented success of deep convolutional neural networks (CNN). More specifically, we investigate U-Net[20], a convolutional encoder-decoder network that has demonstrated exceptional performance in tumor detection and segmentation tasks. Illustrated in Fig. 1a, U-Net takes a three-dimensional (3D) volume image as an input, processes it through a "bottleneck layer" where the image features are compressed, and reconstructs the image into a binary segmentation map indicating a pixel-wise tumor classification result. Here, we focused on the information encoded at the bottleneck layer, which contains rich visual characteristics of the tumor and hypothesized that the encoded information at this layer might be relevant to the tumor malignancy, and thus cancer survival, which is the central hypothesis of this paper.

Results

In prior studies[21,22], we analyzed PET/CT images of 96 NSCLC patients that were obtained within 3 months prior to SBRT, whose summary statistics are illustrated in Extended Data Fig. 1 and Extended Data Table 1. For each volume image, the region of interest (ROI) with a dimension of 96 mm × 96 mm × 48 mm was set around each SBRT treated tumor location and the image was cropped to the ROI volume. Two separate U-Net models were trained to perform tumor segmentation in PET and CT images, respectively. Each of the models was trained with 38 ROI images and the corresponding physician contours, but no other information such as survival time was provided. Segmentation performance was tested on 22 independent ROI images that were not included in training, and the average Sørensen-Dice similarity coefficients (DSC) were 0.861 ± 0.037 (mean ± std) and 0.828 ± 0.087 for CT and PET, respectively. After training, each U-Net model learned to encode 55,296 features at the bottleneck layer for each patient, resulting in a total of 110,592 features per patient.

These features are an intermediate throughput of U-Net, and are then decoded to generate an automated segmentation in the network. It is likely that these features summarize some rich structural and functional geometry of the intratumoral and peritumoral area, some of which might be relevant to cancer survival. We test this proposition by conducting validation studies and examining the statistical prognostic power of those features. One challenge here is that the number of features (d = 110, 592) is substantially larger than the number of cases, making a statistical analysis prone to overfitting. To this end, we first reduce the number of features via an unsupervised feature selection process in which the survival information is hidden. The *k*-medoids clustering method[23] is employed to serve this purpose, as the method is known to be able to select representative features from a large pool of inter-correlated features in similar literature[24]. For the feature similarity measure in *k*-medoids clustering, we use the Pearson correlation distance defined as $1 - R$, where R is the Pearson correlation coefficient. The optimal number of clusters is determined by the Silhouette method[25]. Finally, given the optimal k clusters, the

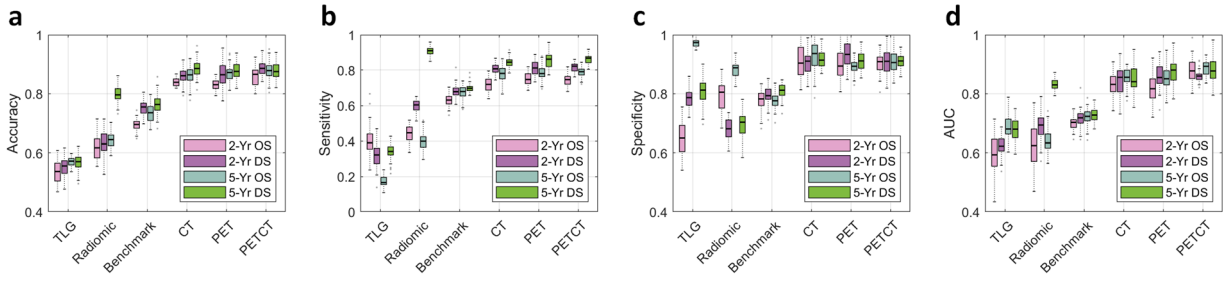

Figure 2. Prognostic performance of the U-Net features. There are four survival categories being tested: 2-year overall survival (2-yr. OS), 5-year overall survival (5-yr. OS), 2-year disease-specific survival (2-yr. DS), and 5-year disease-specific survival (5-yr. DS). The U-Net features are compared against the conventional TLG metric, the 17 radiomics features defined in Oikonomou *et al.*[34] and the benchmark CNN prediction model in Hosny *et al.*[27]. The box plots represent the average performance scores as indicted by the central mark and 25th and 75th percentiles across 6-fold cross validation experiments. (**a**) Overall prediction accuracy (proportion of the correct prediction over the entire data set). (**b**) Sensitivity (correct prediction of death over all death cases). (**c**) Specificity (correct prediction of survival across all survival cases). (**d**) AUC of the receiver operating characteristics (ROC) curve.

medoids of the clusters are selected to be the representative features of the clusters and form an unsupervised feature pool. We choose the final set of features via the least absolute shrinkage and selection operator (LASSO)[26] on the preliminary feature pool obtained from the *k*-medoids method. LASSO attempts to select a few features among the cluster centers that have a strong relationship with the survival outcomes using the fixed regularization parameter. The selected features then serve as regressors in the logistic regression model to predict the cancer survival outcome. We first searched the optimal number of clusters for *k*-medoids clustering via the Silhouette method[25]. Since the performance of *k*-medoids clustering is also dependent on a random initialization of the cluster centers, we tested 10 times for each *k* and computed the mean of the summation of the inner cluster distances to ensure consistency of the clustering outcome. We then computed the curve of the Silhouette values with respect to the number of clusters in total to determine the optimal number of clusters. The medoids of the feature clusters are selected to form a reduced feature set.

We then examined the selected features and their prognostic power via cross-validation on the NSCLC data set collected at the University of Iowa Hospitals and Clinics (UIHC). As summarized in Extended Data Fig. 1 and Extended Data Table 1, the UIHC data set is comprised of primary and follow-up PET/CT images of total 96 NSCLC cases with their survival status and other clinical meta data. Information on the cause of death is also available so that the deceased cases can be further broken down to overall and disease-specific deaths. On this data set, we aim to predict four survival categories, namely, 2-year overall survival (2OS), 5-year overall survival (5OS), 2-year disease-specific survival (2DS), and 5-year disease-specific survival (5DS). Total $N = 96$ cases qualify for 2OS category, 74 for 5OS, 92 for 2DS, and 45 for 5DS, depending on the survival status and the cause of death. Other clinical meta data, such as sex, age, smoking history, and tumor, lymph nodes, and metastasis (TNM) staging, exist in the data set and may provide an improved prognostic power when added as regressors. However, we exclude all other parameters but U-Net encoded image features, in order to focus the analysis on the image features only. Only primary images are encoded via U-Net and used for prediction. Follow-up images are reserved for comparison and discussion later in this paper. Each cross validation experiment is comprised of unsupervised feature selection using *k*-medoids clustering, LASSO-based feature selection, and training of the logistic regression model. These tasks are performed independently from the other cross-validation experiments. In each experiment, we measure the accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) per each survival category. We compute the average and the standard deviation of these performance metrics across the cross-validation experiments to derive the estimated performance metrics and their 95% confidence intervals (95% CI).

In such cross-validation experiments, estimated AUC of the proposed prediction model is 0.88 (95% CI: 0.80–0.96) for the prediction of 2OS. For other survival criteria, the estimated AUCs are similar, namely 0.89 (95% CI: 0.85–0.93) for 5OS, 0.86 (95% CI: 0.81–0.91) for 2DS, and 0.88 (95% CI: 0.81–0.95) for 5DS. Note that the estimated AUC values of the conventional TLG and other radiomics markers[27] range between 0.60 and 0.83 on the same data set. A more recent deep learning method reported in Hosny *et al.*[28] produces AUCs between 0.70∼0.73 on the same data set. A graphical illustration of the result as well as the full set of performance metrics are available in Fig. 2 and Extended Data Table 2, respectively.

Moreover, we further validated the result on an extramural dataset provided by the Stanford Cancer Institute. The Stanford data set is comprised of primary CT images of 26 NSCLC cases which received SBRT, of which 18 survived and 8 died according to 2 year OS and 1 survived and 25 died according to 5 year OS. Neither PET images nor disease-specific survival information were available in the Stanford data set. Training of the U-Net, feature selection (*k*-medoids and LASSO), and construction of the survival model are performed only on the UIHC data set, and the Stanford data set is used for validation only. In this setting, the estimated AUC is 0.87 (95% CI: 0.80–0.94) for 2 year OS, and 0.90 (95% CI: 0.82–0.98) for 5 year OS. More detailed results are illustrated in Fig. 3 and Extended Data Table 3.

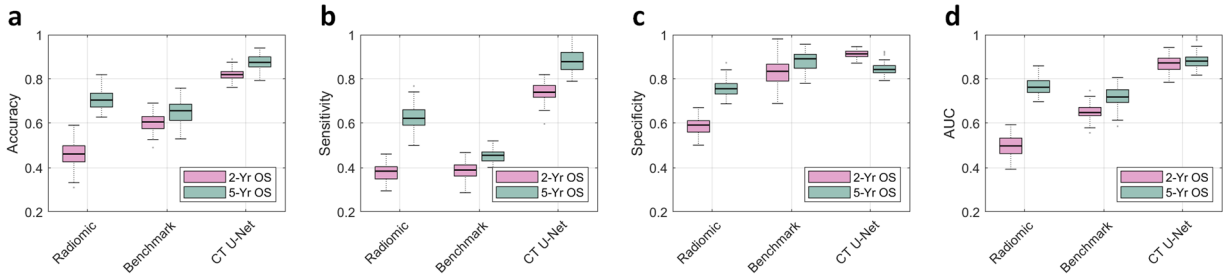

Figure 3. Prognostic performance on an extramural data set. The extramural data set provided by Stanford only includes CT images and not PET images. Additionally, the disease-specific survival information is not provided. Therefore, in this experiment, two survival categories are being tested: 2-year overall survival (2-yr. OS) and 5-year overall survival (5-yr. OS). The U-Net features are compared against the 6 CT-based radiomics features (Radiomic) defined in Oikonomou *et al.*[34] and the benchmark CNN prediction model (Benchmark) in Hosny *et al.*[27]. The box plots represent the average performance scores as indicted by the central mark and 25th and 75th percentiles across experiments tested on extramural Stanford data set. (**a**) Overall prediction accuracy (proportion of the correct prediction over the entire data set). (**b**) Sensitivity (correct prediction of death over all death cases). (**c**) Specificity (correct prediction of survival over all survival cases). (**d**) AUC of the receiver operating characteristics (ROC) curve.

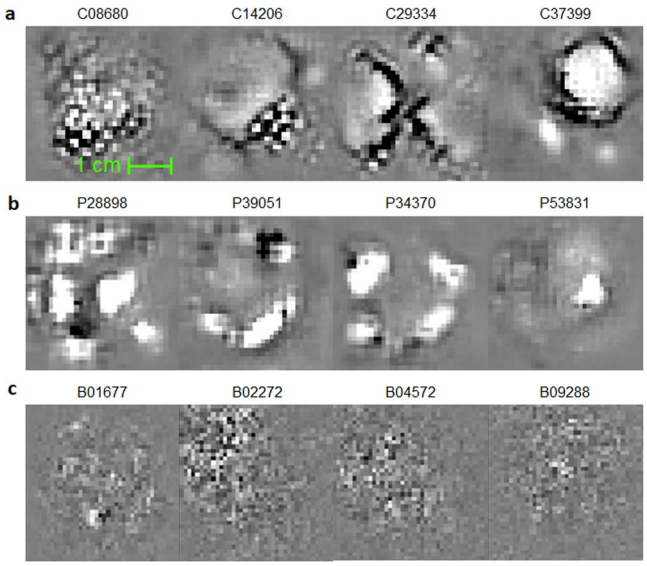

Figure 4. Survival-related features captured by the U-Nets. During training, CNNs essentially learn "templates/patterns" from training images and apply these templates to analyze and understand images. (**a**) Image templates that the U-Nets have captured for the segmentation task, in CT and (**b**) in PET. Note these templates are learned in an unsupervised manner, without any survival-related information provided, despite which these were later discovered to be survival-related. Note that the templates captured by U-Net are characterized by their sensical and interpretable geometric structures. For example, C37399 appears to be a template looking for a tumor-like shape at the top-right corner and a tube-like structure at the bottom-left. In addition, C08680 appears to look for a textural feature of the tumor. (**c**) In contrast, image templates learned by direct fitting of a CNN model to the survival data[27]. Note the features in (**c**) are less interpretable compared to the U-Net features.

Meanwhile, we visualized the features learned by U-Net to develop an intuitive understanding of what those prognostic markers represent. The image features learned by U-Net are essentially artificial neurons in deep neural networks. In principle, we can visualize a neuron by showing multitudes of image patterns and observing which image pattern activates the neuron the most. Practically, we employ an optimization-based approach[29] where the objective is to maximize an individual neuron's activation value by manipulating the input image pattern:

$$\mathbf{X}^* = \underset{\mathrm{X}}{\mathrm{argmax}} q(\mathbf{X}|\mathbf{W}, \mathbf{b}) \tag{1}$$

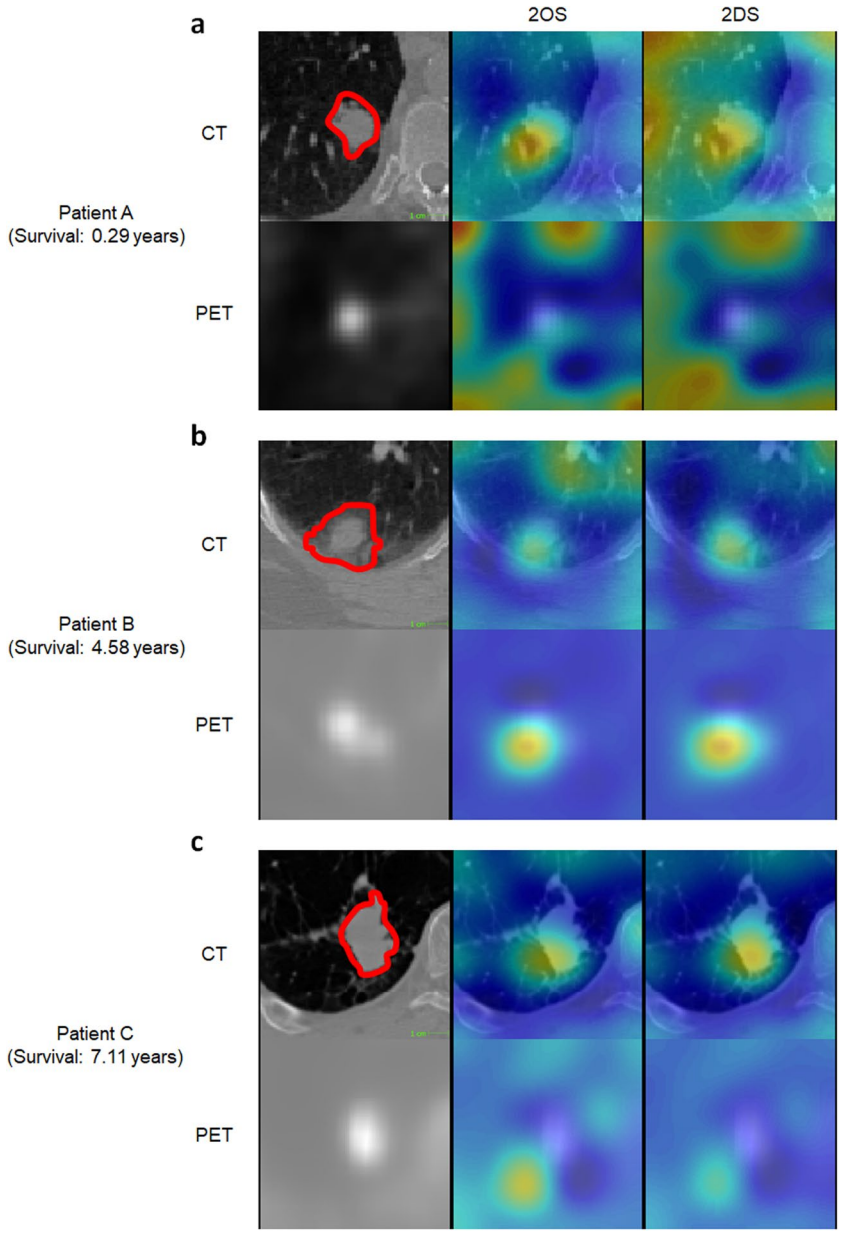

Figure 5. Visualization of the U-Net features. Regions that predicted death of the patients obtained via a guided backpropagation method[29]. Trivially, tumoral regions are highlighted in red in the heatmap. However, some of the heated regions outside of the tumoral volume matched with the actual locations of recurrences and metastases when they were compared with the post-therapeutic images and clinical records, rendering a great potential as a practical, clinical tool for patient-tailored treatment planning in the future. (**a**) Patient deceased in 0.29 years after the acquisition of the images. (**b**) Deceased after 4.58 years. (**c**) Deceased after 7.11 years.

where $q(\cdot|\mathbf{W}, \mathbf{b})$ is the U-Net encoder with the trained model parameters **W** and **b**, and **X** is the input image pattern. Displayed in Fig. 4 and Extended Data Fig. 2 are visualizations of features captured by artificial neurons.

We also visualized which regions in the patient images predicted low survival probability. We employed a guided gradient backpropagation approach[30]. The main idea of the guided backpropagation algorithm is to compute $\frac{\partial P}{\partial x_{i,j,k}}$ where P is the probability of death and $x_{i,j,k}$ is a voxel value at the position (i, j, k) in the patient image. The gradient $\frac{\partial P}{\partial x_{i,j,k}}$ can be interpreted as the change of the death probability when the voxel $x_{i,j,k}$ changes to a different value. If the voxel was not so significant in predicting death, the gradient value would be small, where as if the voxel played an important role in predicting high probability of death, the gradient value would be greater. Displayed in Fig. 5 and Extended Data Figs. 3–5 are heat maps representing the gradient. Heated regions (red) are the areas that lowered the probability of survival whereas the other areas (blue) are the ones that had negligible effect on the survival.

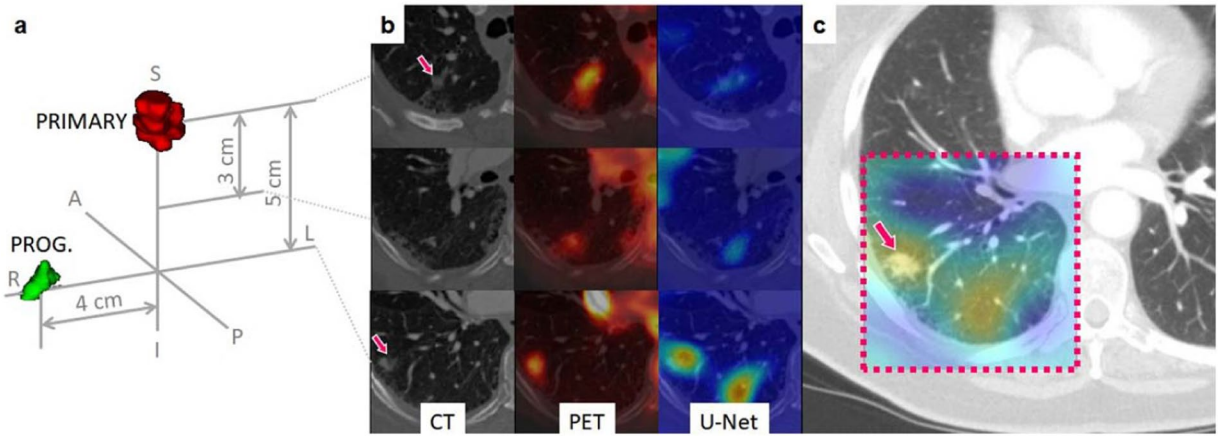

Figure 6. Correlation between U-Net visualization and cancer progression. Post-SBRT CT images were compared with the U-Net visualization results. We observed an agreement of the heated regions with the actual location of recurrence as in this example. (**a**) A 3D rendering showing the location of the primary tumor volume (red) and the progression region (green) of the case IA001765; (**b**) Pre-SBRT transversal slices at the primary tumor location (top), 3 centimeters below (middle row), and 5 centimeters below (inferior row); (**c**) A follow up (post-SBRT) image of the same patient. The dashed box indicates the estimated corresponding ROI to the primary (pre-SBRT) CT slices. The heat map generated based on pre-SBRT, *i.e.*, the same heat map as in the bottom row of (**b**), is superimposed on top of the ROI on the post-SBRT image. Notice that the recurrence location coincides with the heated area.

Discussion

As illustrated Fig. 2 and Extended Data Table 2, the contrast in the predictive performance between the U-Net features and the conventional imaging features (SUV_{MEAN}, SUV_{MAX} and TLG) was evident, quantitatively proving the strong prognostic power of the U-Net features. There also was a noticeable enhancement of prognostic performance when the U-Net features were compared with a recent deep learning approach as in Hosny *et al.*[27]. The same trend could be observed in Fig. 3 and Extended Data Table 3 where we validated the U-Net features against an extramural data set, confirming the enhanced prognostic performance of the U-Net features. Here, it is worth reemphasizing that the U-Net was trained without any survival-related information, and, thus it is highly unlikely that the U-Net-learned features were overfitted to the survival data or biased towards them. Nonetheless, while these U-Net features were identified independently from the survival data, the U-Net features demonstrated strong evidence of correlation and thus prognostic power for NSCLC survival as discussed above.

Here, a justification should be necessary on the rationale behind taking a "detour" by training a segmentation network first, then extracting prognostic image features, and training a survival model, as opposed to directly training a CNN model to the survival data as in other literature[27,31,32]. To this end, we make the following arguments. First, a CNN model trained on a segmentation task is more robust to overfitting and is more generalizable, as the image features themselves are developed from an unsupervised training. A direct prediction of survival from an image tends to be less robust and less generalizable, as the millions of parameters in a CNN can easily misconstrue the trend, and this process is difficult to control. Secondly, segmentation-trained features contain more structural, geometric features that are human-interpretable, whereas direct-trained features tend to be biased towards less-intuitive and fuzzy texture patterns. It is, in fact, well-understood now that CNNs tend to be biased towards texture rather than shape, while the human cognitive process works in the opposite[33]. For example, Geirhos *et al.*[34] conducted an experiment where the texture of a cat in an image was replaced by the texture of an elephant—CNNs started to discern the image as an elephant while humans still categorized the image as a cat. From this, we can infer that the direct training of a CNN for survival prediction would result in a development of less sensical features lacking geometric information. In contrast, segmentation-trained features are incentivized to focus on geometric structures and shapes rather than textures, and thus are more sensical.

The above arguments are indeed supported by the visualization results in Fig. 4 and Extended Data Fig. 2, where the benchmark CNN method appears to be capturing some noisy texture patterns. In contrast, many of the U-Net features appear to be capturing more sensical and interpretable geometric shapes from the image, such as tumor-like blobs (e.g. C00048, C25988, P39051, P47258) or heterogeneity of the tumor (e.g. C08680). Interestingly, some of the U-Net features, for example C01777 and C37399, were looking for tube-like structures nearby the tumor-like blobs, which might be capturing blood vessels and lymphatics in the peritumoral area. This is consistent with the widely accepted clinical knowledge that tumors can show enhanced growth in the presence of nearby vessels and lymphatics as they carry nutrition to supply the tumoral growth.

Meanwhile, we further visualized the proposed prediction model by creating a heat map highlighting regions that predicted low survival probability. The map generated by the guided back-propagation algorithm confirmed that the prediction model was looking at clinically sound regions for making predictions by responding more to tumor and surrounding tumor regions than other regions. In addition to the results presented above, this is yet more evident that the prediction model has produced generalizable features and rules for making prognostic

prediction, and is thus not overfitted. Moreover, through comparison with post-SBRT CT images and clinical records of the patients, we observe that the heat map has the potential to identify regions of progressions or recurrence. For the case illustrated in Fig. 6 as an example, a small nodule had been found initially at the progression region, but was not treated by SBRT as the region was not originally identified as tumor by PET. However, this region was identified as a high risk region on heat map by the U-Net as well as the primary tumor region of SBRT. Tumor recurrence was found on post-SBRT CT images and matched the highlighted risk region on pre-SBRT U-Net generated heat map. From such observations, the heat map visualization has the potential to identify regions at high risk for tumor progression or recurrence that could be utilized for the purpose of assisting patient-tailored treatment planning in the future. For this reason, we believe this indicates more rigorous risk map developments and requires quantitative follow-up validations, which will be our subsequent project.

In summary, we discovered that the U-Net segmentation algorithm trained for automated tumor segmentation on PET/CT, codifies rich structural and functional geometry at the bottleneck layer. These codified features, in turn, could be used for survival prediction in cancer patients even though the U-Net was trained without any survival-related information. The survival model based on such U-Net features demonstrated significantly higher predictive power compared to conventional PET-based, metabolic burden metrics such as TLG or relatively recent hand-crafted radiomics approaches. The validity of this discovery was further confirmed by the validation on an extramural data set provided by the Stanford Cancer Institute. Furthermore, we visualized the survival-related U-Net features and observed that they were indeed depicting intratumoral and/or peritumoral structures that had been previously acknowledged as potentially relevant to cancer survival. Our approach awaits a further validation against a larger number of observations and in a larger variety of cancer types. However, there was not enough clinical evidence to conclude that the visualization of the U-Net features may identify potential regions of recurrence and metastasis and, thus, a follow-up study is suggested. Our findings may suggest a new starting point for quantitative image-based cancer prognosis with a great deal of important new knowledge to be discovered.

Methods

Subjects. PET-CT images of NSCLC patients who received SBRT at the University of Iowa Hospitals and Clinics were investigated in this study. The images were obtained using a dual PET-CT scanner (Biograph 40 PET/CT, Siemens Medical Solutions USA, Inc., Hoffman Estates, IL). CT image used for SBRT planning were co-registered with CT images of PET-CT datasets using a deformable image registration (DIR). Using the deformation map (vectors), PET images were resampled with the primary CT images of SBRT plan. Post-SBRT follow-up images were used only for qualitative validation of the visualization results. The gross tumor volume (GTV) for each of the images was delineated by radiation oncologists on both CT and PET images with the guidance of the corresponding images in the other modality. All tumor contouring was completed using VelocityAI (Varian Medical System, Inc., Palo Alto, CA).

A total of 96 cases (Male = 44; Feamle = 52) were investigated in this study. Patients' group stages vary in sub-categories including 41 in stage I, 10 in stage II, 16 in stage III, 29 in stage IV. Meanwhile, the histology was confirmed by a thoracic pathologist based off visual interrogation of the biopsied or surgically resected specimen. Histologies among the 96 patients include 48 adenocarcinoma, 41 squamous cell, 1 adenosquamous, 3 metastases from previous NSCLC, and 1 without biopsy. For overall survival from SBRT, the 2-year survival rate is 54% and the 5-year survival rate is 6%. For overall survival from diagnosis, the 2-year survival rate is 66% and the 5-year survival rate is 28%. For disease-specific survival from SBRT, the 2-year survival rate is 61% and the 5-year survival rate is 23%. For disease-specific survival from diagnosis, the 2-year disease specific survival rate is 73% and the 5-year disease specific survival rate is 51%. Among a total 96 patients, the qualified portion utilized in this research in each survival sub-category has been visualized and discussed in Extended Data Fig. 1 respectively.

In this study, the total of 96 utilized NSCLC patients were retrospectively analyzed after approval from the University of Iowa Institutional Review Board (IRB: 200503706; Name: The utility of imaging in cancer: staging, assessment of treatment response, and surveillance). All data collection, experimental procedures, and methods applied are in accordance with the relevant guidelines and regulations. All patients consented for the use of their clinical information and medical images. All participants enrolled in this study signed an informed consent developed and approved by the University of Iowa Institutional Review Board. All scans were in digital imaging and communications in medicine (DICOM) format and de-codified.

Data processing. Each pair of the co-registered PET-CT images were resampled with an isotropic spacing in all three dimensions. Each image was then cropped into a fixed size of $96 \times 96 \times 48$ voxels where each voxel represents 1 cubic millimeter (mm^3) after resampling. Within the resampled voxels, intensity values were clipped to the range of $[-500, 200]$ for all CTs and $[0.01, 20]$ for all PETs, to remove outliers. More details are described in our previous work[21,22].

U-net features. In our previous work[22], two independent 3D U-Net networks were constructed and trained for automated tumor segmentation in PET and CT images, respectively. The U-Net networks comprise two major components: the encoder network and the decoder network (Fig. 1). The encoder network takes a $96 \times 96 \times 48$ volume image as an input. The first convolution layer produces 32 features attached to each voxel, representing low-level visual cues such as lines, edges, and blobs. These features are then down-sampled by half in all three dimensions and the down-sampled volume is fed into the second convolution layer, which then produces 64 features per each voxel. This is repeated three more times increasing the number of features to 128, 256, and finally 512, while the volume size is reduced by half in all three dimensions each time. The final $6 \times 6 \times 3 \times 512 (=55,296)$ features that the encoder network produces are an abstract, high-level summary of the input image, which is

then decoded by the symmetric decoder to produce the binary segmentation map (1: tumor, 0: none). The convolutional kernels are of a size $3 \times 3 \times 3$ across all layers and the max-pooling layers of a $2 \times 2 \times 2$ window size with a stride of 2 were used for down-sampling.

To train the PET-CT segmentation U-Net networks, we utilized co-registered PET-CT scan pairs from 60 patients with primary NSCLC. For each PET-CT image, the slice image size is 512×512 and the number of slices varies from 112 to 293. The tumor contour on each of the PET and CT scans were labeled by physicians as groundtruth. In data preprocessing, all PET-CT images are resampled with an isotropic spacing of $1 \times 1 \times 1$ in voxels and then cropped at a fixed size of 3D volumes ($96 \times 96 \times 48$) centered on the mass gravity of each tumor.

All 60 PET-CT scan pairs were split into a training data set with 38 patients and testing data set with 22 patients. Data augmentation was performed using simple translation, rotation and flip operations and the augmented training set has over 3000 3D PET-CT scan pairs respectively.

The 3D-UNet was trained using open source TensorFlow package and ran on NVIDIA GeForce GTX 1080 Ti GPU with 11GB of memory. The Adam optimization method was utilized with a mini-batch size of 1 and for 20 epochs. To prevent overfitting, the weight decay and early-stop techniques were adopted to obtain the best performance on the test set where the DSC value was computed.

After the U-Nets had been trained, PET and CT images of each patient are fed into the U-Nets to produce 55,296 features per imaging modality. These 55,296 features extracted from the U-Net encoder were used for the analyses throughout the paper. The schematic diagram with respect to the methodologies applied in this research for the feature analysis has been illustrated in Fig. 1.

k-medoids feature clustering. In this research, several k-medoids clustering experiments were conducted on the training dataset for the cross-validation experiments. Pearson's correlation distance was employed as the distance metric for clustering the features as expressed in:

$$\mathcal{D}(\boldsymbol{X}, \boldsymbol{Y}) = 1 - \frac{\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\mathrm{Var}(\boldsymbol{X}, \boldsymbol{Y})\mathrm{Var}(\boldsymbol{X}, \boldsymbol{Y})}} \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the two distinct features; $Cov()$ denotes the covariance of the two features and $Var()$ is the variance of the underlying feature. The sum of inner cluster distances were computed by setting various k values, and the optimal number of clusters were determined by the Silhouette method[25]. The medoids of all clusters were selected as candidate features to construct the survival prediction models.

Feature selection. The LASSO regression algorithm was employed to narrow down the scope of analysis to survival-related features from the medoids of clusters obtained from clustering results. The LASSO algorithm used in this study is expressed as:

$$\min_{\beta} \|y - \beta X\|_2^2 + \lambda \|\beta_1\| \tag{3}$$

where y denotes the survival outcome (1: alive, 0: dead), X is a vector containing all latent variables extracted from the U-Net network, β denotes the coefficient of regression, and λ is the penalty coefficient. The L1-norm in the second term penalizes the selection of redundant variables. The parameter λ was determined via cross validation on the training dataset. Latent variables that survived the L1-penalty with the best λ were selected for the logistic regression model to predict the survival outcome. Using the LASSO-selected variables, we applied logistic regression to estimate the coefficients and predict the survival outcome.

Logistic regression. We formulate survival as a dummy variable. The task of predicting survival outcome can then be formulated as a binary class probability prediction problem and we select the linear logistic regression model as our statistical model:

$$y = \{1 + e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}\}^{-1} \tag{4}$$

where y denotes the predicted survival probability, x_i is the i-th LASSO-selected U-Net feature, and β_i is the regression coefficient. The performance of a prediction model was measured via 6-fold cross validation. In experiment, the samples were split into training and test sets. The models were trained with the training set and the test sets were left out for validation. The proportion of survival and death cases were controlled to be equal in the test sets. Reported performance metrics in this paper are based on the statistics of the test set validations across 6-fold cross validation.

Visualization. An activation maximization scheme[28] was employed to visualize the LASSO-selected U-Net features. For a trained U-Net encoder $\mathbf{X} = q(\cdot|\mathbf{W}, \mathbf{b})$, neurons at the bottleneck layer corresponding to the LASSO-selected features were denoted as q_i. Then, Eq. 1 was solved for each individual neuron via gradient ascent:

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \gamma^{(k)} \nabla q_i(\mathbf{X}^{(k)}) \tag{5}$$

where $\mathbf{X}^{(k)}$ is the current solution at k-th iteration and $\gamma^{(k)}$ is a step length. We set $\gamma^{(k)}$ as $1/\sigma^{(k)}$ where $\sigma^{(k)}$ denotes the standard deviation of the gradients. The gradient ∇qi was computed using the standard backpropagation

algorithm. The initial image $\mathbf{X}^{(0)}$ was initialized with random voxel values following the Gaussian distribution $\mathcal{N}$ (128,1). Displayed in Fig. 4a,b are the final solution $\mathbf{X}^{*}$ after 20 iterations.

Moreover, we also visualized a risk map by evaluating each voxel's contribution to the prediction of survival. We employed a guided backpropagation approach similar to Selvaraju *et al.*[29]. For each voxel in the input image, with marginal change of the survival probability with respect to the voxel's intensity, defined as $\frac{\partial P}{\partial x_{i,j,k}}$, where P is the probability of death and $x_{i,j,k}$ is a voxel value at the position (i, j, k). In the guided backpropagation process, we rectified the gradient by dropping the negative gradient values to focus on the "risk". This was achieved by applying rectified linear unit (ReLU) activation when the values were backpropagated from node to node:

$$\alpha^{(m)} = \max\left(\frac{\partial(P)}{\partial \boldsymbol{A}^{(m)}_{i,j,k}}, 0\right) \tag{6}$$

where $A^{(m)}$ denotes the activation map corresponding to the m-th convolutional kernel at the bottleneck encoding. Note that only the LASSO-selected features were involved in the survival model P such that $\frac{\partial(P)}{\partial \boldsymbol{A}^{(m)}_{i,j,k}}$ is zero most of the time. Finally, the risk map R was defined as a linear combination of all activation maps at the bottleneck layer with the coefficients $\alpha^{(m)}$ obtained from the above:

$$\mathcal{R}(\mathbf{X}) = \sum_m \alpha^{(m)} A^{(m)}(\mathbf{X}) \tag{7}$$

Received: 26 June 2019; Accepted: 31 October 2019;
Published online: 21 November 2019

References

1. World Health Organization. Cancer fact sheet. https://www.who.int/news-room/fact-sheets/detail/cancer Accessed: 2019-02-27. (2018).
2. American Cancer Society. Non-small cell lung cancer. https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/what-is-non-small-cell-lung-cancer.html. Accessed: -02-27.2019.
3. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA: A Cancer J. for Clin.* **68**, 7–30, https://doi.org/10.3322/caac.21442 (2018).
4. Woodard, G. A., Jones, K. D. & Jablons, D. M. *Lung Cancer Staging and Prognosis*, 47–75 (Springer International Publishing, Cham, 2016).
5. Berghmans, T. *et al.* Primary tumor standardized uptake value (SUVmax) measured on fluorodeoxyglucose positron emission tomography (FDG-PET) is of prognostic value for survival in non-small cell lung cancer (NSCLC): A systematic review and meta-analysis (MA) by the european lung cancer working party for the IASLC lung cancer staging project. *J. Thorac. Oncol.* **3**, 6–12, https://doi.org/10.1097/JTO.0b013e31815e6d6b (2008).
6. Paesmans, M. *et al.* Primary tumor standardized uptake value measured on fluorodeoxyglucose positron emission tomography is of prognostic value for survival in non-small cell lung cancer: Update of a systematic review and meta- analysis by the European Lung Cancer Working Party for the International Association for the Study of Lung Cancer Staging Project. *J. Thorac. Oncol.* **5**, 612–619, https://doi.org/10.1097/JTO.0b013e3181d0a4f5 (2010).
7. Bollineni, V. R., Widder, J., Pruim, J., Langendijk, J. A. & Wiegman, E. M. Residual 18F-FDG-PET uptake 12 weeks after stereotactic ablative radiotherapy for stage I non-small-cell lung cancer predicts local control. *Int. J. Radiat. Oncol.* **83**, e551–e555, https://doi.org/10.1016/j.ijrobp.2012.01.012 (2012).
8. Burdick, M. J. *et al.* Maximum standardized uptake value from staging FDG-PET/CT does not predict treatment outcome for early-stage non–small-cell lung cancer treated with stereotactic body radiotherapy. *Int. J. Radiat. Oncol.* **78**, 1033–1039, https://doi.org/10.1016/j.ijrobp.2009.09.081 (2010).
9. Agarwal, M., Brahmanday, G., Bajaj, S. K., Ravikrishnan, K. P. & Wong, C.-Y. O. Revisiting the prognostic value of preoperative 18F-fluoro-2-deoxyglucose (18F-FDG) positron emission tomography (PET) in early-stage (I & II) non-small cell lung cancers (NSCLC). *Eur. J. Nucl. Medicine Mol. Imaging* **37**, 691–698, https://doi.org/10.1007/s00259-009-1291-x (2010).
10. Chen, H. H., Chiu, N.-T., Su, W.-C., Guo, H.-R. & Lee, B.-F. Prognostic value of whole-body total lesion glycolysis at pretreatment FDG PET/CT in non-small cell lung cancer. *Radiology* **264**, 559–566 (2012).
11. Zaizen, Y. *et al.* Prognostic significance of total lesion glycolysis in patients with advanced non-small cell lung cancer receiving chemotherapy. *Eur. J. Radiol.* **81**, 4179 – 4184, https://doi.org/10.1016/j.ejrad.2012.07.009 Imaging in Acute Chest Pain. (2012).
12. Mehta, G., Chander, A., Huang, C., Kelly, M. & Fielding, P. Feasibility study of FDG PET/CT-derived primary tumour glycolysis as a prognostic indicator of survival in patients with non-small-cell lung cancer. *Clin. Radiol.* **69**, 268–274, https://doi.org/10.1016/j.crad.2013.10.010 (2014).
13. Chicklore, S. *et al.* Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. *Eur. J. Nucl. Medicine Mol. Imaging* **40**, 133–140, https://doi.org/10.1007/s00259-012-2247-0 (2013).
14. Lee, G. *et al.* Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: State of the art. *Eur. J. Radiol.* **86**, 297–307, https://doi.org/10.1016/j.ejrad.2016.09.005 (2017).
15. Carvalho, S. *et al.* 18F-fluorodeoxyglucose positron-emission tomography (FDG-PET)-radiomics of metastatic lymph nodes and primary tumor in non-small cell lung cancer (NSCLC) – a prospective externally validated study. *PLOS ONE* **13**, 1–16, https://doi.org/10.1371/journal.pone.0192859 (2018).
16. Fried, D. V. *et al.* Stage III non–small cell lung cancer: Prognostic value of FDG PET quantitative imaging features combined with clinical prognostic factors. *Radiology* **278**, 214–222, https://doi.org/10.1148/radiol.2015142920 PMID: 26176655 (2016).
17. Zhang, Y., Oikonomou, A., Wong, A., Haider, M. A. & Khalvati, F. Radiomics-based prognosis analysis for non-small cell lung cancer. *Sci. Reports* **7**, Article number: 46349 (2017).
18. Paul, R. *et al.* Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography* **2**, 388–395, https://doi.org/10.18383/j.tom.2016.00211 (2016).
19. Diamant, A., Avishek Chatterjee, M. V., Shenouda, G. & Seuntjens, J. Deep learning in head & neck cancer outcome prediction. *Sci. Reports* **9**, Article No: 2764, https://doi.org/10.1038/s41598-019-39206-1 (2019).
20. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241 (Springer, 2015).

21. Wu, X., Zhong, Z., Buatti, J. & Bai, J. Multi-scale segmentation using deep graph cuts: Robust lung tumor delineation in MVCBCT. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 514–518 (IEEE, 2018).
22. Zhong, Z. *et al*. Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. *Med. physics* **46**, 619–633, https://doi.org/10.1002/mp.13331 (2019).
23. Park, H.-S. & Jun, C.-H. A simple and fast algorithm for k-medoids clustering. *Expert. systems with applications* **36**, 3336–3341 (2009).
24. Uthoff, J. *et al*. Machine learning approach for distinguishing malignant and benign lung nodules utilizing standardized perinodular parenchymal features from CT. *Med. Phys*. (2019).
25. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. computational applied mathematics* **20**, 53–65 (1987).
26. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Ser. B (Methodological)* 267–288 (1996).
27. Oikonomou, A. *et al*. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Sci. reports* **8**, 4003, https://doi.org/10.1038/s41598-018-22357-y (2018).
28. Hosny, A. *et al*. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS medicine* **15**, e1002711 (2018).
29. Yosinski, J., Clune, J., Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. In *ICML Workshop on Deep Learning* (2015).
30. Selvaraju, R. R. *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626, https://doi.org/10.1109/ICCV.2017.74 (2017).
31. Paul, R. *et al*. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography* **2**, 388 (2016).
32. Yao, J., Wang, S., Zhu, X. & Huang, J. Imaging biomarker discovery for lung cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 649–657 (Springer, 2016).
33. Kubilius, J., Bracci, S. & de Beeck, H. P. O. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology* **12**, e1004896 (2016).
34. Geirhos, R. *et al*. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2018).

Acknowledgements

The authors want to express their gratitude to Dr. Ruijiang Li at Stanford University for his assistance in coordinating the collaboration with Stanford University and for providing the extramural dataset. They also thank Dr. Sanjay Aneja at Yale University for providing insight and expertise that greatly assisted the research. Research reported in this publication was supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under award number 1R21CA209874 and partially by U01CA140206 and P30CA086862.

Author contributions

Y.K. and X.W. conceived and managed the experiments, S.B., Y.H. and L.T. conducted and analyzed the experiments, Y.H. and Z.S. performed visualization experiments, S.B. and B.J.S. advised the statistical analysis, Y.H., M.G., and K.R.C. processed the raw data, B.G.A., J.M.B., and K.A.P. provided the contours for training of the U-net, B.G.A., J.M.B., and S.N.S. advised clinical discussions. R.L., J.W., M.D., and B.L. provided and processed the extramural data set. All authors reviewed the manuscript.

Competing interests

Dr. Buatti's work has been funded in part by the National Institutes of Health/National Cancer Institute grants P01 CA217797001A1, UL1 TR002537, U01 CA140206, and 1R21CA209874. Dr. Kim's work has been funded in part by the National Institutes of Health/National Cancer Institute grant 1R21CA209874. Dr. Wu's work has been funded in part by the National Institutes of Health grants 1R21CA209874 and R01EB020665. Drs. Baek, Wu, Kim, Smith, Allen, Buatti and Mr. He are the co-inventors of the provisional patent (U.S. App. No. 62/811,326. Systems And Methods For Image Segmentation And Survival Prediction Using Convolutional Neural Networks) based upon the work of this manuscript. Dr. Plichta, Dr. Seyedin, Ms. Gannon, and Ms. Cabel declare no potential conflict of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-53461-2.

Correspondence and requests for materials should be addressed to Y.K. or X.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.