Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

Precision Medicine and Imaging

Clinical Cancer Research

Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging

Yiwen Xu[1], Ahmed Hosny[1,2], Roman Zeleznik[1,2], Chintan Parmar[1], Thibaud Coroller[1], Idalid Franco[1], Raymond H. Mak[1], and Hugo J.W.L. Aerts[1,2,3]

Abstract

Purpose: Tumors are continuously evolving biological systems, and medical imaging is uniquely positioned to monitor changes throughout treatment. Although qualitatively tracking lesions over space and time may be trivial, the development of clinically relevant, automated radiomics methods that incorporate serial imaging data is far more challenging. In this study, we evaluated deep learning networks for predicting clinical outcomes through analyzing time series CT images of patients with locally advanced non–small cell lung cancer (NSCLC).

Experimental Design: Dataset A consists of 179 patients with stage III NSCLC treated with definitive chemoradiation, with pretreatment and posttreatment CT images at 1, 3, and 6 months follow-up (581 scans). Models were developed using transfer learning of convolutional neural networks (CNN) with recurrent neural networks (RNN), using single seed-point tumor localization. Pathologic response validation was performed on dataset B, comprising 89 patients with NSCLC treated with chemoradiation and surgery (178 scans).

Results: Deep learning models using time series scans were significantly predictive of survival and cancer-specific outcomes (progression, distant metastases, and local-regional recurrence). Model performance was enhanced with each additional follow-up scan into the CNN model (e.g., 2-year overall survival: AUC = 0.74, $P < 0.05$). The models stratified patients into low and high mortality risk groups, which were significantly associated with overall survival [HR = 6.16; 95% confidence interval (CI), 2.17–17.44; $P < 0.001$]. The model also significantly predicted pathologic response in dataset B ($P = 0.016$).

Conclusions: We demonstrate that deep learning can integrate imaging scans at multiple timepoints to improve clinical outcome predictions. AI-based noninvasive radiomics biomarkers can have a significant impact in the clinic given their low cost and minimal requirements for human input.

Introduction

Lung cancer is one of the most common cancers worldwide and the highest contributor to cancer death in both the developed and developing worlds (1). Among these patients, most are diagnosed with non–small cell lung cancer (NSCLC) and have a 5-year survival rate of only 18% (1, 2). Despite recent advancements in medicine spurring a large increase in overall cancer survival rates, this improvement is less consequential in lung cancer, as most symptomatic and diagnosed patients have late-stage disease (3). These late-stage lesions are often treated with nonsurgical approaches, including radiation, chemotherapy, targeted, or immunotherapies. This signals the dire need for monitoring therapy response using follow up imaging and tracking radiographic changes of tumors over time (4). Clinical response assessment criteria, such as RECIST (5), analyze time series data using simple size-based measures such as axial diameter of lesions.

Artificial intelligence (AI) allows for a quantitative, instead of a qualitative, assessment of radiographic tumor characteristics, a process also referred to as "radiomics" (6). Indeed, several studies have demonstrated the ability to noninvasively describe tumor phenotypes with more predictive power than routine clinical measures (7–10). Traditional machine learning techniques involved the derivation of engineered features for quantitative description of images with success in detecting biomarkers for response assessment and clinical outcome prediction (11–15). Recent advancements in deep learning (6) have demonstrated successful applications in image analysis without human feature definition (16). The use of convolutional neural networks (CNN) allows for the automated extraction of imaging features and identification of nonlinear relationships in complex data. CNN networks that have been trained on millions of photographic images can be applied to medical images through transfer learning (17). This has been demonstrated in cancer research with regards to tumor detection and staging (18). AI developments can be clinically applicable to enhance patient care by providing accurate and efficient decision support (6, 11).

The majority of quantitative imaging studies have focused on the development of imaging biomarkers for a single timepoint (19, 20). However, the tumor is a dynamic biological system with vascular and stem cell contributions, which may respond, thus the phenotype may not be completely captured at a single timepoint (21, 22). It may be beneficial to incorporate posttreatment

[1]Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts. [2]Radiology and Nuclear Medicine, GROW, Maastricht University Medical Centre, Maastricht, the Netherlands. [3]Department of Radiology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Hugo J.W.L. Aerts, Harvard–Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115. Phone: 617-525-7156; Fax: 617-525-7156; E-mail: hugo_aerts@dfci.harvard.edu

doi: 10.1158/1078-0432.CCR-18-2495

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

Translational Relevance

Medical imaging provides noninvasive means for tracking patients' tumor response and progression after treatment. However, quantitative assessment through manual measurements is tedious, time-consuming, and prone to interoperator variability, as visual evaluation can be nonobjective and biased. Artificial intelligence (AI) can perform automated quantification of radiographic characteristics of tumor phenotypes as well as monitor changes in tumors before, during, and after treatment in a quantitative manner. In this study, we demonstrated the ability of deep learning networks to predict prognostic endpoints of patients treated with radiation therapy using serial CT imaging routinely obtained during follow-up. We also highlight their potential in accounting for and utilizing the available serial images to extract the relevant timepoint and image features pertinent to the prediction of survival and response to treatment. This provides further insight into applications including the detection of gross residual disease without surgical intervention, as well as other personalized medicine practices.

CT scans from routine clinical follow-up as a means to tracking changes in phenotypic characteristics after radiation therapy. State of the art deep learning methods in video classification and natural language processing have utilized recurrent neural networks (RNN) to incorporate longitudinal data (23). However, only a few studies have applied these advanced computational approaches in radiology (24).

In this study, we use AI in the form of deep learning, specifically CNNs and RNNs, to predict survival and other clinical endpoints of patients with NSCLC by incorporating pretreatment and follow up CT images. Two datasets were analyzed containing patients with similar diagnosis of stage III lung cancer, but treated with different therapy regimens. In the first dataset, we developed and evaluated deep learning models in patients treated with definitive chemoradiation therapy. The generalizability and further pathologic validation of the network was evaluated on a second dataset comprising patients treated with chemoradiation followed by surgery. For localization of the tumors, only single-click seed points were needed without volumetric segmentations, demonstrating the ease of incorporating a large number of scans at several timepoints into deep learning analyses. The CT imaging-based patient survival predictions can be applied to response assessment in clinical trials, precision medicine practices, and tailored clinical therapy. This work has implications for the use of AI-based imaging biomarkers in the clinic, as they can be applied noninvasively, repeatedly, at low cost, and requiring minimal human input.

Materials and Methods

Patient cohorts

We used two independent cohorts, dataset A and dataset B, consisting in a total of 268 patients with stage III NSCLC for this analysis. Dataset A contained 179 consecutive patients who were treated at Brigham and Women's/Dana-Farber Cancer Center between 2003 and 2014 with definitive radiation therapy and chemotherapy with carboplatin/paclitaxel or cisplatin/etoposide (chemoRT) and had at least one follow-up CT scan. We analyzed a total of 581 CT scans (average of 3.2; range 2–4 scans per patient, 125 attenuation CTs from PET and 456 diagnostic CTs) of pretreatment and follow-up scans at 1, 3, and 6 months after radiation therapy for delta analysis of the serial scans (Fig. 1). The CT–PET scans were acquired without iodinated contrast, and the contrast administration of chest CT scans are patient specific and based on clinical guidelines. As a realistic representation of clinical settings, not all patients received imaging scans at all timepoints (Supplementary Fig. S1). Patients with surgery prior to or after therapy were not included in this study. The main endpoint of this study was the prediction of survival and prognostic factors for stage III patients treated with definitive radiation (Fig. 2). Dataset A was randomly split 2:1 into training/tuning ($n = 107$) and test ($n = 72$). Overall survival was assessed along with three other clinical endpoints for the definitive radiation therapy cohort: distant metastases, locoregional recurrence, and progression.

An additional test was performed on dataset B, a cohort of 89 consecutive patients with stage III NSCLC from our institution between 2001 and 2013, who were treated with neoadjuvant radiotherapy and chemotherapy prior to surgical resection (trimodality). The analysis of dataset B was included for further validation with a range of standard of care treatment protocols. A total of 178 CT scans with two timepoints; scans taken prior to radiation therapy and the scans after radiation were used, both taken prior to surgery. Patient exclusion included those who presented with distant metastasis or those with more than a 120 days delay between chemoradiation and surgery, as well as those without survival data. For both cohorts, no histologic exclusions were applied. The endpoint of the additional test set of trimodality patients was the prediction of pathologic response, validated at the time of surgery. The residual tumor was classified as responders (pathologic complete response $n = 14$, and microscopic residual disease $n = 28$) or gross residual disease ($n = 47$) based on surgical pathologic reports.

CT acquisition and image preprocessing

CTs were acquired according to standardized scanning protocols at our institution, using a GE "Lightspeed" CT scanner (GE Medical System) for treatment, pretreatment, and follow-up scans. The follow-up scans consisted of different axial spacing and a portion of the images are from PET–CT acquisitions. The input of the tumor image region is defined at the center of the identified seed point for the pretreatment, and for the 1, 3, and 6-month follow-up CT scans after definitive radiation therapy. The seed points were manually defined in 3D Slicer 4.8.1 (25). Because of the variability in slice thicknesses and in-plane resolution, the CT voxels were interpolated to $1 \times 1 \times 1$ mm^3 using linear and nearest neighbor interpolation. To have a stable input for the proposed architecture, it was necessary to interpolate the imaging data to homogeneous resolution. This was performed as the slice thicknesses were a maximum of 5 mm and thus the 2D input images are taken at a slice not further than 2 mm away from a non-interpolated slice. The linear interpolation was used to avoid potential perturbations from more complex interpolation methods, which involves and may be dependent on several parameters and longer computation time. The fine scale was chosen to maintain the details of the tumor.

Three axial slices of 50×50 mm^2 centered on the selected seed point were used as inputs to the model. They were spaced 5 mm

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

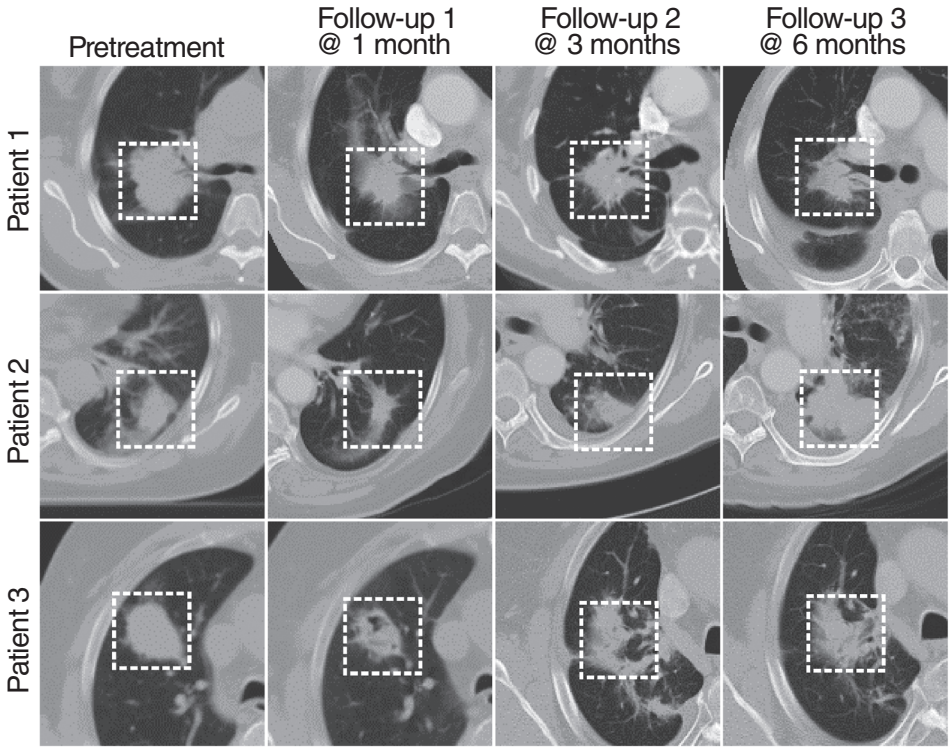

Figure 1.
Serial patient scans. Representative CT images of patients with stage III nonsurgical NSCLC before radiation therapy and 1, 3, and 6 months following radiation therapy. A single click seed point identifies the input image patch of the neural network (defined by the dotted white line).

apart; the center slice is on the same axial slice as the seed point. 5 mm was the maximum slice thickness of the CT images. A transfer learning approach was applied using the pretrained ResNet CNN that was trained on natural RGB images. The three axial slices were used as input to the CNN network. Using three 2D slices gives the network information to learn from but keeps the number of features lower than a full 3D approach, reduces GPU memory usage and training time, as well as limits the overfitting. Image augmentation was performed on the training data, and involved image flipping, translation, rotation, and deformation, which is a conventional good practice and has shown to improve performance (26). The same augmentation was performed on the pretreatment and follow-up images, such that the network generates a mapping for the entire input series of images. The deformation was on the order of millimeters and did not noticeably change the morphology of the tumor or surrounding tissues.

Neural network structure

The network structure was implemented in Python, using Keras with Tensorflow backend (Python 2.7, Keras 2.0.8, Tensorflow 1.3.0). The proposed network structure has a base ResNet CNN trained on the ImageNet database containing over 14 million natural images (Fig. 3). One CNN was defined for each timepoint input, such that an input with scans at three timepoints would involve input into three CNNs. The output of the pretrained network model was then input into recurrent layers with gated recurrent units (GRU), which takes the time domain into account. To ensure the network was able to handle missing scans (27, 28),

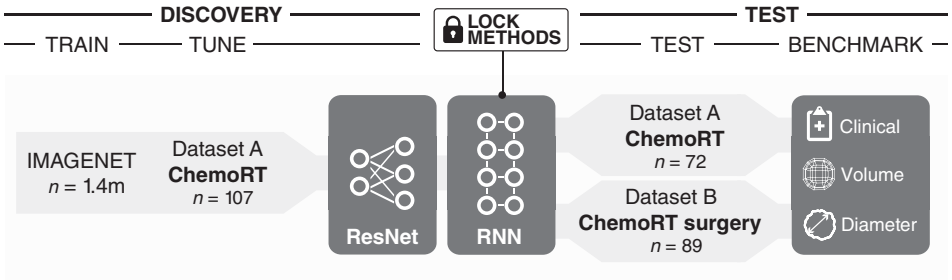

Figure 2.
Analysis design. Depiction of the deep learning–based workflow with two datasets and additional comparative models. Dataset A included patients treated with chemotherapy and definitive radiation therapy, and was used to train and fine-tune a ResNet CNN combined with an RNN for predictions of survival. A separate test set from this cohort was used to assess performance and compared with the performance of radiographic and clinical features. Dataset B included patients treated with chemotherapy and surgery. This cohort was used as an additional test set to predict pathologic response, and the model predictions were compared with the change in volume.

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

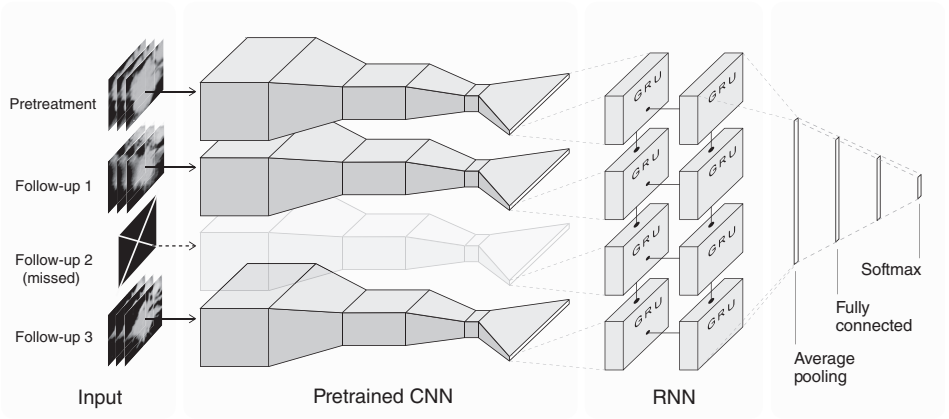

Figure 3.

Deep learning architectures. The neural architecture includes ResNet CNNs merged with an RNN, and was trained on baseline and follow-up scans. The input axial slices of 50 × 50 mm^2 centered on the selected seed point were used as inputs to the model. They were spaced 5 mm apart; the center slice is on the same axial slice as the seed point. Deep learning networks are trained on natural RGB images and thus need three image slices for input. The outputs of each CNN model are input into the RNN, with a GRU for time-varying inputs. Masking was performed on certain inputs of the CNN so that the recurrent network takes missed scans into account. The final softmax layer provides the prediction.

RNN algorithms were used which allowed for amalgamation of several timepoints and the ability to learn from samples with missed patient scans at a certain timepoints. The output of the pretrained network was masked to skip the timepoint when a scan was not available. Averaging and fully connected layers are then applied after the GRU with batch normalization (29) and dropout (30) after each fully connected layer to prevent overfitting. The final softmax layer allows for a binary classification output. To test a model without the input of follow-up scans, the pretreatment image alone was input into the proposed model, with the recurrent and average pooling layers replaced by a fully connected layer, as there was only one input timepoint.

Transfer learning

Weights trained with ImageNet (26, 31), a set of 14 million 2D color images, were used for the ResNet (31) CNN and the additional weights following the CNN were randomized at initialization for transfer learning. Dataset A was randomly split 2:1 into training/tuning and test. Training was performed with Monte Carlo cross-validation, using 10 different splits (further 3:2 split of training:tuning) on 107 patients with class weight balancing for up to 300 epochs. The model was evaluated on an independent test set of 72 patients, who were not used in the training process. The surviving fractions for training/tuning ($n = 107$) and test sets ($n = 72$) were comparable (Supplementary Table S1). Only the pretreatment image was input into the proposed model, and the recurrent and average pooling layers were replaced with a fully connected layer.

Statistical analysis

Statistical analyses were performed in Python version 2.7. All predictions were evaluated on the independent test set of dataset A for survival and for prognostic factors after definitive radiation therapy. The clinical endpoints included distant metastasis, progression, and locoregional recurrence as well as overall survival for 1 and 2 years following radiation therapy. The analyses were compared with a random forest clinical model with features of stage, gender, age, tumor grade, performance, smoking status, and clinical tumor size (primary maximum axial diameter).

Statistical differences between positive and negative survival groups in dataset A were assessed using the area under the receiver operator characteristic curve (AUC), and the Wilcoxon rank sums test (also known as the Mann–Whitney *U* test). Prognostic and survival estimates were calculated using the Kaplan–Meier method between low and high mortality risk groups, stratified at the median prediction probability of the training set and controlled using a log-rank test. Hazard ratios were calculated through the Cox proportional-hazards model.

An additional test was performed on dataset B, the trimodality cohort using the 1-year survival model from the definitive radiation cohort with two timepoints. Survival predictions were made from the 1-year survival model trained on dataset A. The model predictions were used to stratify the trimodality patients based on survival and tumor response to radiation therapy prior to surgery. The groups were assessed using their respective AUC, and were tested with the Wilcoxon rank sums test. This was compared with the volume change after radiation therapy and a random forest clinical model with the same features used for dataset A.

Results

Clinical characteristics

To evaluate the value of deep learning based biomarkers to predict overall survival using patient images prior and post radiation therapy (Fig. 1), a total of 268 patients with stage III NSCLC with 739 CT scans were analyzed (Fig. 2). Dataset A consisted of 179 patients treated with definitive radiation therapy and was used as a cohort to train and test deep learning biomarkers (Supplementary Table S2). There was no significant difference between the patient parameters in the training and test sets of dataset A ($P > 0.1$, group summary values in Supplementary Table S2). The patients were 52.8% females (median age of 63 years; age range 32–93 years) and were predominantly diagnosed as having stage IIIA (58.9%) NSCLC at the time of diagnosis, with 58.1% in the adenocarcinoma histology category. The median radiation

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

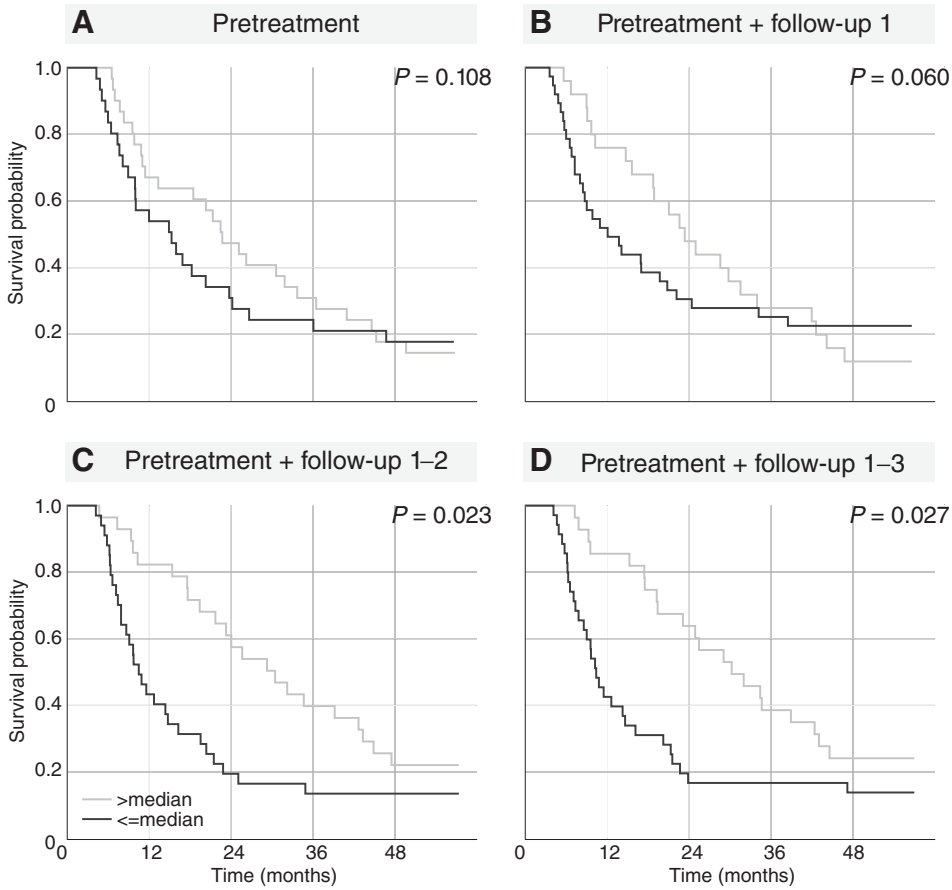

Figure 4.
Performance deep learning biomarkers on validation datasets. The deep learning models were evaluated on an independent test set for performance. The 2-year overall survival Kaplan–Meier curves were performed with median stratification (derived from the training set) of the low and high mortality risk groups with no follow-up or up to three follow-ups at 1, 3, and 6 months posttreatment for dataset A (72 definitive patients in the independent test set, log-rank test $P < 0.05$ for > one follow-up).

dose was 66 Gy for the definitive radiation cohort (range 45–70 Gy, median follow-up of 31.4 months). Another cohort of 89 patients treated with trimodality served as an external test set (dataset B). The median radiation dose for the trimodality patients was lower, at 54 Gy (range 50 to 70 Gy, median follow-up of 37.1 months).

Deep learning–based prognostic biomarker development and evaluation

To develop deep learning–based biomarkers for overall survival, distant metastasis, disease progression, and locoregional recurrence, training was performed using the discovery part of dataset A (Fig 2). To leverage the information from millions of photographic images, the ResNet CNN model was pretrained on ImageNet and then applied to our dataset using transfer learning. The CNN extracted features of the CT images of each timepoint were fed into a recurrent network for longitudinal analysis. We observed that baseline model with only pretreatment scans demonstrated low performance for predicting 2-year overall survival (AUC = 0.58; $P = 0.3$; Wilcoxon test). Improved performance to predict 2-year overall survival was observed with the addition of each follow-up scan; at 1 month (AUC = 0.64, $P = 0.04$), 3 months (AUC = 0.69, $P = 0.007$), and 6 months (AUC = 0.74, $P = 0.001$; Supplementary Fig. S2). We also observed the similar trend in performance for other clinical endpoints, that is 1-year, survival, metastasis, progression, and locoregional recurrence-free survival (Supplementary Fig. S3). A clinical model, incorporating stage, gender, age, tumor grade, performance, smoking status, and clinical tumor size, did not yield a statistically significant prediction of survival (2-year survival AUC = 0.51, $P = 0.93$) or treatment response (Supplementary Table S3).

Further survival analyses were performed with Kaplan–Meier estimates for low and high mortality risk groups based on median stratification of patient prediction scores (Fig. 4). The models for 2-year overall survival yielded significant differences between the groups with two ($P = 0.023$, log-rank test) and three ($P = 0.027$, log-rank test) follow-up scans. Comparable results were found for the following predictions with their respective hazard ratios: 1-year overall survival (6.16; 95% CI, 2.17–17.44]; $P = 0.0004$), distant metastasis free (3.99; 95% CI, 1.31–12.13; $P = 0.01$), progression free (3.20; 95% CI, 1.16–8.87; $P = 0.02$), and no locoregional recurrence (2.74; 95% CI, 1.18–6.34; $P = 0.02$), each with significant differences at three follow-up timepoint scans.

Predicting pathologic response

As an additional independent validation and to evaluate the relationship between delta imaging analysis and pathologic response, the trimodality pre-radiation therapy and post-radiation therapy prior to surgery scans were input into the neural network model trained on dataset A. First for survival prediction evaluation, the model was tested on dataset B. To match the number of input timepoints, the 1-year survival model with the pretreatment and first follow-up at 1 month was used. The model significantly predicted distant metastasis, progression, and local regional recurrence (Supplementary Table S4). Although, for

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

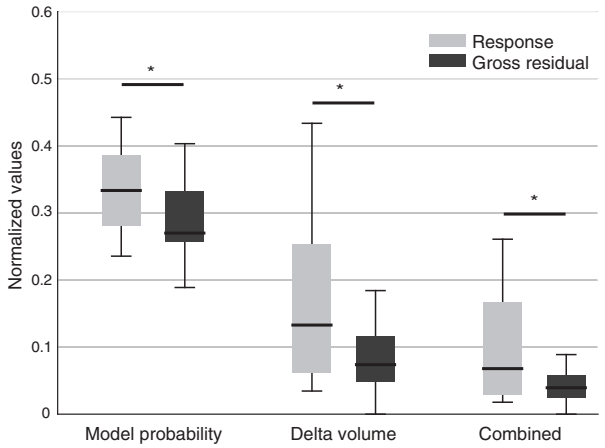

Figure 5.

Pathologic response prediction validation. Model probability and the change in volume after radiation therapy was used for the prediction of pathologic response. The CNN survival model significantly stratified response and gross residual disease in the second test set dataset B; comparable predictions were found with change in tumor volume and the combination of the two parameters ($n = 89$; Wilcoxon, $P < 0.05$).

overall survival there were a low number of events (30 of 89), the model was trending towards making a prediction for 3-year overall survival in dataset B.

The predictions of the network were then used to categorize pathologic response (Fig. 5), and were found to significantly distinguish between responders and gross residual disease, with an AUC of 0.65 ($n = 89$; $P = 0.016$; Wilcoxon test), which was similar to the change in volume (AUC of 0.65; $n = 89$; $P = 0.017$; Wilcoxon test). To investigate the additive performance, we build a combined model of the network probabilities and change in volume, which showed slightly higher performance (AUC of 0.67; $n = 89$; $P = 0.006$; Wilcoxon test). The CNN probabilities and changes in the primary tumor volume were significantly correlated ($P = 0.0002$), although with a Spearman's correlation value of 0.39. A clinical model, involving parameters of stage, gender, age, tumor grade, performance, smoking status, and clinical tumor size, did not yield a statistically significant prediction for pathologic response ($P = 0.42$; Wilcoxon test).

Discussion

Tracking tumor evolution for prediction of survival and response after chemotherapy and radiation therapy can be critical to treatment assessment and adaptive treatment planning for improving patient outcomes. Conventionally, clinical parameters are used to determine treatment type and to predict outcome (2), but this does not take into account phenotypic changes in the tumor. Medical imaging tracks this evolution of lesions noninvasively and provides a method for tracking the same region longitudinally through time, providing additional tumor characteristics beyond those obtained through static images at a single timepoint (5). Follow-up CT scans are already a part of the clinical workflow, providing additional information regarding the patient. Using deep learning approaches for tumor assessment allows for the extraction of phenotypic changes without manual and/or semiautomated contours or qualitative visual interpretations, which are prone to interobserver variability. Additionally, prognostic predictions can potentially aid in the assessment of patient outcome in clinical trials to assess response and eventually dynamically adapting therapy.

Using a combined image-based CNN and a time encompassing RNN, the neural network was able to make survival and prognostic predictions at 1 and 2 years for overall survival. As expected, with an increase in the number of timepoints and the amount of imaging data available to the network, there was an increase in performance. Although the performance varied between the predictions, there was a consistent increase in AUC, due to the increase in signal from each additional image of the primary tumor and the changes between the scans with time. In this cohort, using a single pretreatment scan was not successful in making a prediction of survival. However, previous work in the field of radiomics using engineered (9, 12, 14, 15) and deep learning (10) approaches using pretreatment imaging data only, were able to predict the endpoint of their interest with the use of anatomical CT or functional PET data. For the cohorts in this study, there is a trend towards significance of the deep learning model with the pretreatment timepoint only. Using larger cohorts could improve the predictive power of the imaging markers. The clinical model, which included the clinical tumor size (longest axial diameter), was also not predictive of survival or the other prognostic factors.

The neural network was able to stratify patients into low and high mortality risk groups, with significant difference in overall survival (Fig. 4). This was also identified for the risk of locoregional recurrence with the input of two follow-up timepoints at around 1 and 3 months after the completion of definitive radiation therapy. The other outcomes, progression, and distant metastasis needed the additional third follow-up at around 6 months for a significant stratification of the mortality risk groups. This may be due to a more defined set of early imaging phenotypes relating to survival and locoregional recurrence as compared with the other prognostic factors, or confounding phenotypes with regards to distant metastasis and progression, which the model cannot overcome unless the third follow-up is incorporated.

The two datasets within our study are inherently different as the cohorts are comprised of patients with different disease burdens and treatment modalities. The surgical patients are younger and healthier on average, with an earlier stage of disease, and well enough to tolerate surgery. It has been shown that the survival of surgical patients is dependent on the success of the surgical procedure and distant disease (32), where definitive radiation therapy survival is determined by local control (33). There was also a higher proportion of stage IIIA in patients who also underwent surgical resection (dataset B) compared with definitive radiation therapy patients (dataset A).

Despite these differences, the survival CNN models trained on dataset A predicted surrogates of survival in dataset B including distant metastasis, progression, and locoregional recurrence. It was trending towards predicting survival and this may be due to the inherent differences between the cohorts, as well as the low number of events in the cohort and sample size. There was also only one follow-up scan available for dataset B, thus less information was provided to the survival model. Although the model was designed to overcome the immortal time bias, there could still be an effect. With more timepoints, fewer patients are alive to have

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

the scan performed and thus decrease the ability to predict survival.

Survival is associated with tumor pathologic response (34, 35). Thus, we tested the relationship between the probabilities of the survival network model on similar patients with stage III NSCLC who were in different treatment cohorts (definite radiation therapy and trimodality). Dataset B included the follow-up timepoint after radiation therapy and prior to surgery, for the prediction of response and for further validation of our model. This also serves as a test for generalizability in locally advanced NSCLC patients treated with different standard of care treatment protocols. To match the number of input timepoints, the 1-year overall survival model with the pretreatment and first follow-up at 1 month was used. The model was able to separate the pathologic responders from those with gross residual disease in the trimodality cohort. This was the case, even though the model development was completely blinded from this cohort.

This prediction was compared with a well-known prediction of response, the primary tumor size. The change in tumor volume also predicted the response in this cohort with a similar performance. However, the two measures, model probability and delta volume, were only weakly correlated and the combined model showed a slight improvement in performance. The proposed model was able to predict pathologic response in a different cohort, with only the image and a seed point for input. There is also a weak correlation between the values, which suggests that the image-based neural network model is detecting radiographic characteristics other than tumor size.

The use of a CNN-based network captures the tumor region and the immediate tumor environment. Previous techniques focused on providing the machine learning algorithm with accurate manual delineations or semiautomated methods, which may not incorporate surrounding tissue (36, 37). CNN image input includes the boundary between the tumor and the normal tissue environment. This may provide additional indications for tumor response and infiltration to the surrounding tissue. Image augmentation was performed on the training tumor region, as conventional practice in the field of deep learning and biomedical image processing (38), to improve performance and the small-scale deformations were applied to prevent overfitting (39) on our relatively small training set. The use of conventional ResNet CNN for image characterization allows for the incorporation of pretreatment weights on natural images (26). This mediated the application of deep neural networks on medical images, with cohorts much smaller than the millions of samples used in other AI solutions.

The number of samples available for most radiologic studies are not on the same order of magnitude as those used for deep learning applications. For instance, a facial recognition deep learning application was developed by training on 87,000 images and testing on 5,000 images (40). However, transfer learning can be used to leverage common low-level CNN parameters from databases such as ImageNet, which contains over 14 million natural images (26). It would be ideal incorporate the whole tumor volume by using a network pretrained on 3D radiographic images or 3D images in general, however the number of images available are not near the order of magnitude of which are in photographic images. If available, a model pretrained in 3D CT images with samples on the order of thousands of images will likely be overfitted to the patient cohort, the institution, and the outcome the network was trained to predict. The use of transfer learning has demonstrated its effectiveness on improving the performance of lung nodule detection in CT images (18). Our study contained a sample size not on the order of studies based on photographic images, but the current performance was made possible with the incorporation of pretrained networks on ImageNet. Transfer learning may also be used to test the feasibility of clinically applicable utilities prior to the collection of a full cohort for analysis.

The incorporation of follow-up timepoints to capture dynamic tumor changes was key to the prediction of survival and tumor prognosis. This was feasible with the use of RNNs, which allowed for amalgamation of several timepoints and the ability to learn from samples with missed patient scans at a certain timepoint, which is inevitable in retrospective studies such as this one. Although this type of network has not been applied to medical images, similar network architectures have demonstrated success in image and time-dependent analyses, as in video classification and description applications (41). The model was structured to overcome the immortal time bias (42). The pooling of CNN without the RNN has been previously applied (43), but in this case would result in bias classifications for an event when the last patient scan is missed. The RNN was set to not learn from inputs where there is a missing scan (44). GRU RNNs were used as they contain an update gate and a reset gate, which decides the weighting of the information passed on to the network output (45). This captures the pertinent information from each timepoint for the survival and prognostic predictions.

Previous work has demonstrated the feasibility of using CT imaging features to make associations and predictions in lung cancer (7). Several studies used radiomics approaches involving manual delineation of the tumor volume and user defined calculated features to make predictions of survival and pathologic response (12–15). Recent applications of deep learning on lung cancer has focused on lung nodule classification as benign or metastatic and they focus on a single scan for the model input. The study by Kumar and colleagues depended on manual delineation of lung nodules with feature extraction using an autoencoder and classification with decision trees (46). Hua and colleagues used 2D region of the tumor lesion on the axial slice for classification, also performed at one timepoint (47). Our study differs mainly in the incorporation of multiple timepoints in the prediction of survival and prognostic factors. For further validation, we also applied our developed model on a different cohort for the prediction of pathologic response, an important clinical factor. In comparison to previous studies, our model only takes a seed point and creates a 50×50 mm^2 region around the seed point, which is used as input. To compute handcrafted radiomic features, an accurate tumor delineation is required (9), which is susceptible to inter-reader segmentation variability and also is time-consuming. Recently, deep learning has been shown to have higher performance than conventional radiomics (39). Our approach only required a seed point within a tumor and hence is more efficient and robust to manual inference. Additional clinical and pathologic evaluations are not always available. Morphologic parameters dependent on manual and semiautomated contours of the whole tumor volume or RECIST (5) measurements are prone to interoperator variability and can be costly to acquire.

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

Ideally, after training on a larger diverse population and after extensive external validation and benchmarking with current clinical standards, quantitative prognostic prediction models can be implemented in the clinic (48). There are several lung nodule detection algorithms available in the literature and with the aid of the pretreatment tumor contours routinely delineated by the radiation oncologist, the location of the tumor on the follow up images can be detected automatically (49). The input of our model would simply be the bounding box surrounding the detected tumor and can be cropped automatically as well. The trained network can generate probabilities of prognosis within a few seconds, and thus would not hinder current clinical efficiency. The probabilities can then be presented to the physician along with other clinical images and measures, such as the RECIST criteria (5), to aid in the process of patient assessment.

This proof of principle study has its limitations, one of which is the sample size of the study cohorts. Thus, a pretrained CNN was used to improve predictive power. Using a deep learning technique has its limitations. Previous associations were found for risk of distant metastases with the pretreatment scan only, with machine learning techniques (15). It has been demonstrated that machine learning based on engineered features out performs deep learning with small sample sizes. Perhaps with a larger cohort, we could potentially achieve better performance deep learning. The probabilities are essentially calculated with a black box for a specific task, thus are less practical than engineered features, which could potentially be reused for other applications. Neural networks can be prone to overfitting, even with the techniques we have used to mitigate this (29, 30), thus images were resampled to a common pixel spacing. Our model used three 2D slices due to the predefined parameters necessary for transfer learning. However, a 3D image volume may better represent tumor biology and thus increase performance. Our survival models are based purely on the CT image and could potentially benefit from the incorporation of patient specific parameters, such as age, sex, histology, smoking cessation, and radiation therapy parameters, with a larger cohort of patients. With these limitations, our deep learning model was able to make predictions of survival and perhaps with a larger dataset and finer more consistent axial spacing, higher and more clinically relevant performance may be feasible.

Deep learning is a flexible technique which has been successfully implemented in several fields (16). However, the theory behind how the network functions has yet to be established (50). The input and output of the model can be quite intuitive, but as suggested by the term, the hidden middle layers are not. It is therefore very challenging to determine the reasoning behind a network's performance and whether certain parameters have a positive or negative impact. Unlike engineered features built to capture certain characteristics of the image, the interpretation of deep learning features can be ambiguous. To circumvent this in the field of image-based CNN, activation maps have been generated to capture highly weighted portions of the image with respect to the network's predictions (65). This can be visualized in the form of heat maps, generated over the final convolutional layer. Also, how to incorporate the domain knowledge into these abstract features is a very important question that needs to be addressed. Further research in this direction could make these automatically learned feature representations more interpretable.

Conclusions

This study demonstrated the impact of deep learning on tumor phenotype tracking before and after definitive radiation therapy through pretreatment and CT follow-up scans. There were increases in performance of survival and prognosis prediction with incorporation of additional timepoints using CNN and RNN networks. This was compared with the performance of clinical factors, which were not significant. The survival neural network model could predict pathologic response in a separate cohort with trimodality treatment after radiation therapy. Although the input of this model consisted of a single seed point input at the center of the lesion, without the need for volumetric segmentation our model had comparable predictive power compared with tumor volume, acquired through time-consuming manual contours. Noninvasive tracking of the tumor phenotype predicted survival, prognosis, and pathologic response, which can have potential clinical implications on adaptive and personalized therapy.

Disclosure of Potential Conflicts of Interest

R.H. Mak is a consultant/advisory board member for AstraZeneca, Varian Medical Systems, and NewRT. H.J.W.L. Aerts holds ownership interest (including patents) in Sphera and Genospace. No potential conflicts of interest were disclosed by the other authors.

Disclaimer

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

Conception and design: Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, R.H. Mak, H.J.W.L. Aerts
Development of methodology: Y. Xu, A. Hosny, R. Zeleznik, H.J.W.L. Aerts
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): Y. Xu, T. Coroller, R.H. Mak, H.J.W.L. Aerts
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, R.H. Mak, H.J.W.L. Aerts
Writing, review, and/or revision of the manuscript: Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R.H. Mak, H.J.W.L. Aerts
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Y. Xu, R. Zeleznik, T. Coroller, I. Franco, H.J.W.L. Aerts
Study supervision: Y. Xu, H.J.W.L. Aerts

Acknowledgments

The authors acknowledge financial support from the NIH (NIH-USA U24CA194354, and NIH-USA U01CA190234); https://grants.nih.gov/funding/index.htm.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 15, 2018; revised December 19, 2018; accepted January 28, 2019; published first April 22, 2019.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65:87–108.
2. Ettinger DS, Akerley W, Borghaei H, Chang AC, Cheney RT, Chirieac LR, et al. Non-small cell lung cancer. J Natl Compr Canc Netw 2012;10:1236–71.

3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin 2016;66:7–30.
4. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WEE, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM Classification for Lung Cancer. J Thorac Oncol 2016;11:39–51.
5. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228–47.
6. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Hugo J W. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10.
7. Parmar C, Grossmann P, Bussink J, Lambin P, Hugo J W. Machine learning methods for quantitative radiomic biomarkers. Sci Rep 2015;5:13087.
8. Aerts HJWL. Data science in radiology: a path forward. Clin Cancer Res 2018;24:532–4.
9. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.
10. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLoS Med 2018;15:e1002711.
11. Parmar C, Barry JD, Hosny A, Quackenbush J, Aerts HJ. Data analysis strategies in medical imaging. Clin Cancer Res 2018;24:3492–9.
12. Coroller TP, Agrawal V, Huynh E, Narayan V, Lee SW, Mak RH, et al. Radiomic-based pathological response prediction from primary tumors and lymph nodes in NSCLC. J Thorac Oncol 2017;12:467–76.
13. Huynh E, Coroller TP, Narayan V, Agrawal V, Hou Y, Romano J, et al. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. Radiother Oncol 2016;120:258–66.
14. Coroller TP, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee SW, et al. Radiomic phenotype features predict pathological response in non-small cell lung cancer. Radiother Oncol 2016;119:480–6.
15. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RTH, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114:345–50.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
17. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–98.
18. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–98.
19. Dandil E, Cakiroglu M, Eksi Z, Ozkan M, Kurt OK, Canan A. Artificial neural network-based classification system for lung nodules on computed tomography scans. In: 6th International Conference on Soft Computing and Pattern Recognition; 2014 Aug 11–14; Tunis, Tunisia. Washington (DC): IEEE Computer Society; 2014. p. 382–6.
20. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 2016;35:1285–98.
21. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the evolution of non-small-cell lung cancer. N Engl J Med 2017;376:2109–21.
22. Hermann PC, Huber SL, Herrler T, Aicher A, Ellwart JW, Guba M, et al. Distinct populations of cancer stem cells determine tumor growth and metastatic activity in human pancreatic cancer. Cell Stem Cell 2007;1:313–23.
23. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 2017;39:677–91.
24. Cierniak R. A new approach to image reconstruction from projections using a recurrent neural network. Int J Appl Math Comput Sci 2008;18:147–57.
25. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30:1323–41.
26. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84–90.
27. Rubins J, Unger M, Colice GL. Follow-up and surveillance of the lung cancer patient following curative intent therapy. Chest 2007;132:355S–367S.
28. Calman L, Beaver K, Hind D, Lorigan P, Roberts C, Lloyd-Jones M. Survival benefits from follow-up of patients with lung cancer: a systematic review and meta-analysis. J Thorac Oncol 2011;6:1993–2004.
29. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv Preprint 2015. arxiv.org/abs/1502.03167.
30. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings: 29th IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016; 2016 Jun 26–Jul 1; Las Vegas, NV. Washington (DC): IEEE Computer Society; 2016. p. 770–8.
32. Albain KS, Swann RS, Rusch VR, Turrisi AT, Shepherd FA, Smith CJ, et al. Phase III study of concurrent chemotherapy and radiotherapy (CT/RT) vs CT/RT followed by surgical resection for stage IIIA(pN2) non-small cell lung cancer (NSCLC): outcomes update of North American Intergroup 0139 (RTOG 9309). J Clin Oncol 23, 2005 (suppl; abstr 7014).
33. Tsujino K, Hirota S, Endo M, Obayashi K, Kotani Y, Satouchi M, et al. Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer. Int J Radiat Oncol Biol Phys 2003;55:110–5.
34. Hellmann MD, Chaft JE, William WN Jr, Rusch V, Pisters KMW, Kalhor N, et al. Pathological response after neoadjuvant chemotherapy in resectable non-small-cell lung cancers: proposal for the use of major pathological response as a surrogate endpoint. Lancet Oncol 2014;15:e42–50.
35. Pataer A, Kalhor N, Correa AM, Raso MG, Erasmus JJ, Kim ES, et al. Histopathologic response criteria predict survival of patients with resected lung cancer after neoadjuvant chemotherapy. J Thorac Oncol 2012;7:825–32.
36. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9:e102107.
37. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. Invest Radiol 2015;50:757–65.
38. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III. Berlin: Springer; 2015. p. 234–41.
39. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
40. Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH. Washington (DC): IEEE Computer Society; 2014. p. 1891–8.
41. Joe Yue-Hei Ng, Ng JY-H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, et al. Beyond short snippets: deep networks for video classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. http://dx.doi.org/10.1109/cvpr.2015.7299101.
42. Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167:492–9.
43. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun

Published OnlineFirst April 22, 2019; DOI: 10.1158/1078-0432.CCR-18-2495

23–28; Columbus, OH. Washington (DC): IEEE Computer Society; 2014. p. 1725–32.
44. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018;8:6085.
45. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25–29; Doha, Qatar. Stroudsburg (PA): Association for Computational Linguistics; 2014. p. 1724–34.
46. Kumar D, Wong A, Clausi DA. Lung nodule classification using deep features in CT images. In: Proceedings: 2015 12th Conference on Computer and Robot Vision. CRV 2015; 2015 Jun 3–5; Halifax, Nova Scotia, Canada. Washington (DC): IEEE Computer Society; 2015. p. 133–8.
47. Hua K-L, Hsu C-H, Hidayati SC, Cheng W-H, Chen Y-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 2015;8:2015–22.
48. Lehman CD, Yala A, Schuster T, Dontchos B, Bahl M, Swanson K, et al. Mammographic breast density assessment using deep learning: clinical implementation. Radiology 2018;180694.
49. Valente IRS, Cortez PC, Neto EC, Soares JM, de Albuquerque VHC, Tavares JMRS. Automatic 3D pulmonary nodule detection in CT images: a survey. Comput Methods Programs Biomed 2016;124:91–107.
50. Wang G. A perspective on deep imaging. IEEE Access 2016;4:8914–24.

AACR American Association for Cancer Research

Clinical Cancer Research

Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging

Yiwen Xu, Ahmed Hosny, Roman Zeleznik, et al.

Clin Cancer Res Published OnlineFirst April 22, 2019.