An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction

Bin Lou, Semihcan Doken, Tingliang Zhuang, Danielle Wingerter, Mishka Gidwani, Nilesh Mistry, Lance Ladic, Ali Kamen, Mohamed E Abazeed

Summary

Background Radiotherapy continues to be delivered without consideration of individual tumour characteristics. To advance towards more precise treatments in radiotherapy, we queried the lung CT-derived feature space to identify radiation sensitivity parameters that can predict treatment failure and hence guide the individualisation of radiotherapy dose.

Methods An institutional review board-approved study (IRB 14-562) was used to identify patients treated with lung stereotactic body radiotherapy. Patients with primary (stage IA–IV) or recurrent lung cancer and patients with other cancer types with solitary metastases or oligometastases to the lung were included. Patients without digitally accessible CT image or radiotherapy structure data were excluded. The internal study cohort received treatment at the main campus of the Cleveland Clinic (Cleveland, OH, USA). The independent validation cohort received treatment at seven affiliate regional or national sites. We input pre-therapy lung CT images into Deep Profiler, a multi-task deep neural network that has radiomics incorporated into the training process, to generate an image fingerprint that predicts time-to-event treatment outcomes and approximates classical radiomic features. We validated our findings with the independent study cohort. Deep Profiler was combined with clinical variables to derive *i*Gray, an individualised dose that estimates treatment failure probability to be below 5%.

Findings A total of 1275 patients were assessed for eligibility and 944 met our eligibility criteria; 849 were in the internal study cohort and 95 were in the independent validation cohort. Radiation treatments in patients with high Deep Profiler scores failed at a significantly higher rate than in patients with low scores; 3-year cumulative incidence of local failure in the internal study cohort was 20·3% (16·0–24·9) in patients with high Deep Profiler scores and 5·7% (95% CI 3·5–8·8) in patients with low Deep Profiler scores (hazard ratio [HR]=3·64 [95% CI 2·19–6·05], p<0·0001). Deep Profiler independently predicted local failure (HR=1·65 [1·02–2·66], p=0·042). Models that included Deep Profiler and clinical variables predicted treatment failures with a concordance index (C-index) of 0·72 (95% CI 0·67–0·77), a significant improvement compared with classical radiomics (p<0·0001) or clinical variables (p<0·0001) alone. Deep Profiler performed well in the independent validation cohort, predicting treatment failures across diverse clinical settings and CT scanner types (C-index 0·77, 95% CI 0·69–0·92). *i*Gray had a wide dose range (21·1–277 Gy) and suggested dose reductions in 23·3% of patients. Our results also showed that *i*Gray can be safely delivered in the majority of cases.

Interpretation Our results indicate that there are image-distinct subpopulations that have differential sensitivity to radiotherapy. The image-based deep learning framework proposed herein is the first opportunity to use medical images to individualise radiotherapy dose. Our results signify a new roadmap for deep learning-guided predictions and treatment guidance in the image-replete and highly standardised discipline of radiation oncology.

Funding Siemens Medical Systems USA.

Copyright © 2019 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

Lancet Digital Health 2019; 1: e136–47

See **Comment** page e106

Digital Technology and Innovation Division, Siemens Healthineers, Princeton, NJ, USA (B Lou PhD, L Ladic PhD, A Kamen PhD); **Diagnostic Imaging Computed Tomography, Siemens Healthineers, Malvern, PA, USA** (N Mistry PhD); **and Department of Translational Hematology Oncology Research** (T Zhuang PhD, S Doken BA, D Wingerter BE, M Gidwani BS, M E Abazeed MD) **and Department of Radiation Oncology** (M E Abazeed MD), **Cleveland Clinic, Cleveland, OH, USA**

Correspondence to:
Dr Mohamed E Abazeed, Cleveland Clinic, Cleveland, OH 44195, USA
abazeem@ccf.org

Introduction

Medical imaging is integral to the management of patients with cancer, with roles extending from diagnosis to treatment response monitoring.[1] Its ubiquity in clinical practice notwithstanding, its current use remains largely subjective, exemplified by annotations of dimensions delimited by a range of human exactness. CT is the most frequently used modality across all cancers and comprises information beyond tumour geometry.[2] Information acquired by the scanner is conveyed by a matrix of voxels across thin sections of the body composed of x-ray attenuation values proportional to the density of the incident matter. These values can have a total range of more than 4096 intensities. The human eye, on the other hand, resolves a minor proportion of these intensities.[3,4] Such limited discriminatory capacity calls for machine-like methods of information extraction and knowledge optimisation.

Research in context

Evidence before this study

CT images comprise voxel intensities that, to the extent that they are discernable by the treating physician, can guide the manual delineation of tumour volumes. They do not, however, currently contribute to the individualisation of radiation dose prescriptions. Extraction and analysis of rigidly defined radiomic features can transform medical imaging data into quantifiable variables to predict survival, other failure modes, and response to therapeutic agents. To advance toward more precise treatments in radiotherapy, we queried the lung CT-derived feature space to identify radiation sensitivity parameters that can predict treatment failure and hence guide the individualisation of radiotherapy dose.

Added value of this study

To our knowledge, this study is the first to implement a deep neural network, Deep Profiler, using deformable multitasking with the ability to create new radiomic features to predict the risk of failure for patients treated with radiotherapy. This study also represents an innovation in personalised medicine by the projection of an optimised radiation dose, iGray. The cohort of patients evaluated represents one of the largest datasets of chest CT images used in outcome prediction analysis.

Implications of all the available evidence

Accurate estimates of the likelihood of response to treatments coupled with optimised dose delivery can improve clinical outcomes and reduce toxicity for patients receiving radiotherapy. Our framework could provide readily implementable guidance for treatment strategies in under-resourced medical facilities and populations. The ability of the neural network, Deep Profiler, to generate new predictive features represents an advance in radiomics and artificial intelligence. Augmenting this impact, Deep Profiler's prediction accuracy is scalable because it could improve as our dataset increases via natural growth, federated datasets, and data partitions into more homogeneous populations.

Recent advances in image analysis[5] have allowed for precisely this task. Radiomics permits the extraction of quantitative imaging descriptors or features that could characterise more objective tumour characteristics beyond human detection. This approach converts image data into a high-dimensional feature space using a large number of data-characterisation algorithms.[6,7] Some of the features extracted with radiomics have been shown to capture distinct tumour characteristics and exhibit prognostic power;[8] limitations to handcrafted image features, however, are their manual labelling and their inability to conform to a specific task. Manual labelling confines the feature space to elements that humans can learn, and lack of deformability is a characteristic of the a priori design of the features, which cannot be modified based on the classification task at hand.

The process of feeding a machine raw data, like CT pixels, and allowing it to discover vectors for classification through the use of multiple layers of features is known as deep learning.[9] Compared with natural images, medical images have regulated quality that can reduce noise, making them more useful for approaches based on deep learning.[10] However, although medical images can be an ideal source for deep learning, it remains difficult to secure a large quantity of clinically annotated datasets.[11] Since classification accuracy is dependent on the size of the initial training datasets, computational methods that seek to optimise model performance are crucial.

Cancers are characterised by substantial diversity and the optimal therapeutic approach has been shown to vary on the basis of the genetic features of individual cancers.[12,13] Similarly, image-based profiling of tumours might reveal subpopulations that are more or less likely to be sensitive to particular therapies, which might help in the delivery of these therapies. Classifications made by deep learning algorithms have begun to stratify patients on the basis of the type of cancer and genetic alterations;[14,15] however, very little progress has been made in the use of deep learning to predict tumour responses to individual anticancer therapies.

High-dose radiation delivery to the lung via stereotactic body radiotherapy was developed with the intent to achieve local tumour control while potentially avoiding perioperative or long-term surgical morbidity in patients with early-stage lung cancer or oligometastatic disease to the lung. Despite several prospective clinical trials showing excellent local tumour control in medically inoperable patients with lung cancer,[16–18] recent studies describe unacceptably high rates of local failure in some patient subgroups.[19–21] Ongoing and future studies of lung stereotactic body radiotherapy are likely to be informed by a more accurate and quantitative determination of risk of treatment failure, and the mitigation of that failure by adjustment of radiotherapy dose.

In this study, we input pre-therapy lung CT images into Deep Profiler, a multitask deep neural network that has radiomics incorporated into the training process. We combined these data with clinical variables to derive *i*Gray, an individualised radiation dose that results in an estimation of failure probability below 5% at 24 months.

Methods

Study design and participants

An institutional review board-approved study (IRB 14-562) was used to identify patients treated with lung stereotactic body radiotherapy. From a list of 1275 patients treated with lung stereotactic body radiotherapy at our umbrella institution (Cleveland Clinic main and other sites), we

included patients who had evaluable electronic health record data and CT images. Patients were not included if either they were immediately lost to follow-up (0 months) or they did not have an archived image available to us for evaluation. Patients with primary (stage IA–IV) or recurrent lung cancer and patients with other cancer types with solitary metastases or oligometastases to the lung were included. Patients without digitally accessible CT image or radiotherapy structure data were excluded. The internal study cohort received treatment at the main campus of the Cleveland Clinic (Cleveland, OH, USA). The independent validation cohort received treatment at the seven affiliate regional or national sites (Fairview Hospital, Hillcrest Hospital, Independence Family Health Center, Cancer Center Mansfield, North Coast Cancer Center, Wooster Specialty Center, and Cleveland Clinic Florida).

Patients were treated on the basis of either a pathological or radiographical diagnosis. Primary lung cancers were staged using chest CTs. PET and imaging of the brain (MRI or CT) was used when clinically indicated. In cases where imaging revealed mediastinal or hilar lymph nodes enlarged by accepted radiographic criteria or where the standardised uptake value was above 3·0 on PET, pathological mediastinal evaluation with endobronchial ultrasonography-guided sampling was requested.

Patients were immobilised with abdominal compression to restrict breathing motion before receiving radiotherapy. In cases where motion could not be adequately restricted to less than 1 cm, Active Breathing Coordinator (Elekta, Stockholm, Sweden) was used. Tumours within a 2 cm expansion of the tracheobronchial tree were categorised as central. A risk-adapted approach for radiation dose delivery was used. Most patients received 50 Gy in five fractions. When the Radiation Therapy Oncology Group (RTOG) 0236 trial[16] started, eligible patients with peripheral tumours (those enrolled in the trial[16]) received up to 60 Gy in three fractions. Patients with central tumours continued to receive 50 Gy in five fractions. Alternative fractionations were used for patients enrolled in a clinical trial (other than RTOG 0236)[16] at the discretion of the treating radiation oncologist, or if constraints for our standard fractionation schedules could not be met. Local failure was defined as radiographical progression within 1 cm of the planning target volume, to be consistent with definitions of local or marginal failure in clinical trials of stereotactic body radiotherapy.[16] Failures within the same lobe of the lung but greater than 1 cm from the planning target volume of the initial treatment site were defined as lobar failure and were not considered in this analysis. 8·5% of patients received adjuvant chemotherapy. The main indication for patients to receive adjuvant chemotherapy was a perceived high risk of treatment failure. The recommendation to deliver adjuvant treatments was also dependent on considerations of patient tolerance for additional therapy.

Procedures

Planning CT images with corresponding physician-designated gross tumour volumes were used for analysis. Images with contrast were excluded. Four scanners were used for the internal study cohort—namely, three Philips Brilliance CT Big Bore (annotated CT-1, CT-2 and CT-3) and a Philips AcQSim (CT-4). Four scanners were used for the independent validation cohort—namely, GE Medical Systems Discovery ST, Philips Brilliance CT Big Bore, Philips Gemini GXL, and a Siemens SOMATOM Definition AS.

The schema for deriving the Deep Profiler signature is shown in figure 1. A step-by-step protocol for generating Deep Profiler scores and a detailed description of the multi-task learning framework is in the appendix. In brief, Deep Profiler consists of three main parts: an encoder for extracting imaging features and building a task-specific fingerprint, a decoder for estimating handcrafted radiomic features, and a task-specific network for generating image signature for therapy outcome prediction. A three-dimensional (3D) convolutional neural network was used as an encoder for extracting imaging features. We assessed the predictive performance of the deep neural network based on a nested five-fold cross-validation experiment, where 20% of the dataset was used for testing with no overlapping between holdout sets across folds. Within each fold, we used 80% for training the model and 20% for validation. Parameters were optimised using the training set and selected based on the performance in the validation set. The final performance was evaluated using the holdout set. Pre-therapy CT images were first input into Deep Profiler to generate an image signature for treatment outcome prognosis. This signature was then combined with clinical variables in a multivariable model for predicting local failures and estimating *i*Gray.

See Online for appendix

The 3D handcrafted radiomic features were extracted from gross tumour volumes encompassing regions of interest. The handcrafted features can be divided into four groups: (1) intensity, (2) geometry, (3) texture, and (4) wavelet features. The intensity features quantified the first-order statistical distribution of the voxel intensities within the gross tumour volumes. The geometry features quantified 3D shape characteristics of the tumour. The texture features described spatial distribution of the voxel intensities, thereby quantifying the intratumoral heterogeneity. The intensity and texture features were also computed after applying wavelet transformations to the original image. A total of 365 radiomic features were extracted. A list of all features can be found in the appendix. All handcrafted features were extracted using Pyradiomics.[22]

We examined the performance of handcrafted radiomics to predict local failure. A nested five-fold cross-validation experiment was also used for this analysis. Given that the number of radiomic features is much larger than the number of failures, either strong feature

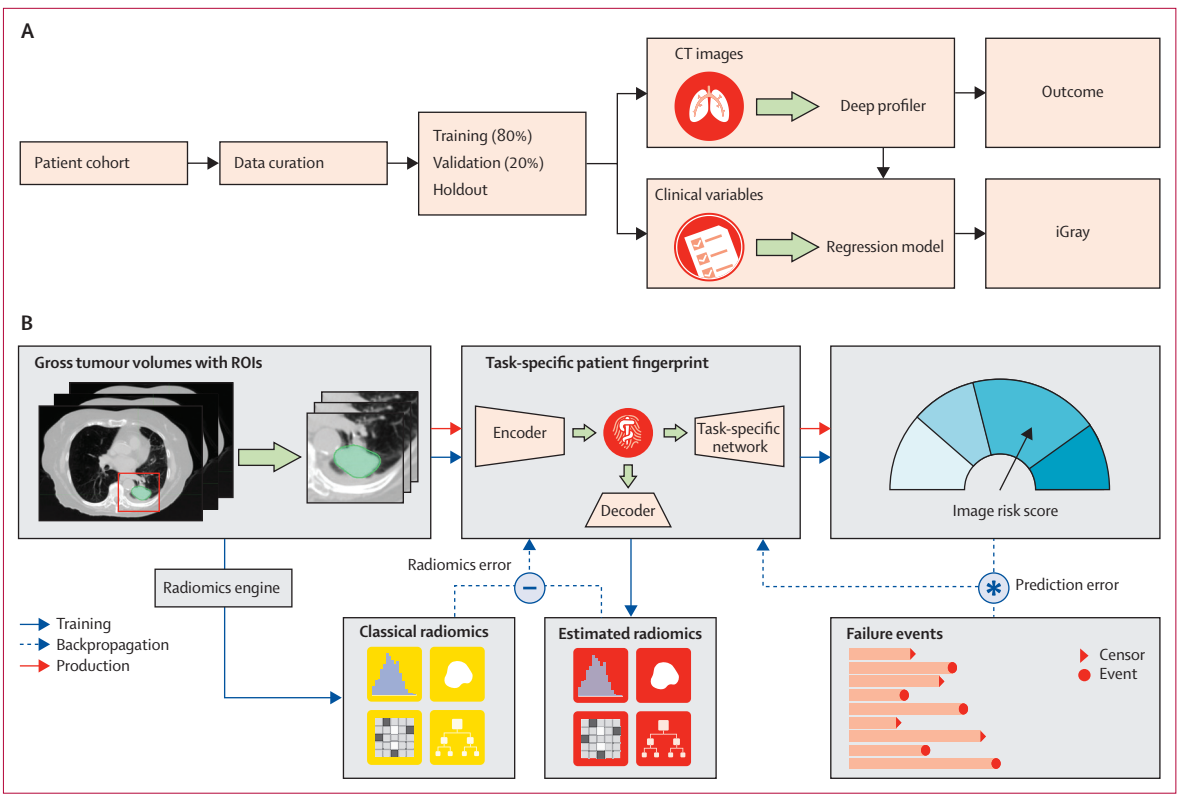

Figure 1: **Study design and neural network architecture**
(A) The predictive performance of Deep Profiler was assessed on the basis of a nested five-fold cross-validation experiment, where 20% of the dataset was used for testing with no overlap between testing sets (holdout sets) across folds. Within each fold, 80% was used for training the model and 20% for validation. The holdout subset provides a final estimate of the model's performance after it has been trained and validated. The model has not seen these data (withheld). *i*Gray was calculated by combining image and clinical data and using the cumulative incidence function of the regression model. (B) Deep Profiler is built of three main parts: an encoder for extracting imaging features and building a task-specific fingerprint, a decoder for estimating handcrafted radiomic features, and a task-specific network for generating an image signature for therapy outcome prediction. The encoder consists of a three-dimensional convolutional neural network. Fully connected networks were adopted in both decoder and task-specific network, aiming to link the latent fingerprint space with classical radiomics and predict the outcome. Since therapeutic outcomes contain time-to-event information, we used a proportional hazards model to relate the time that passes before local failure occurs to classify failures. The failure time could be from an actual failure event or censored at follow-up or death. The blue arrows indicate the training mode, where classical radiomics are calculated using predefined formulas; estimated radiomics and risk scores are the outputs of Deep Profiler model. The dashed lines indicate the backpropagation process. The minus sign means the calculation of radiomics error (ie, reconstruction loss function), which is the difference between estimated and classical radiomics. The asterisk sign means the calculation of prediction error (ie, partial likehihood loss function) based risk scores and event time. Two error terms are combined and backpropagated to optimise the model parameters. A detailed description of loss functions is provided in the appendix. The red arrows show the production mode, where only risk scores are generated by the model. ROI=region of interest.

selection or model regularisation was required to prevent overfitting. Feature selection was done as previously described.[8] In the training set, we computed the performance of all individual features using the concordance index (C-index) and selected the one best feature from each of the four feature groups. These four features were then combined in a multivariable model for predicting local failure. Parameters estimated from the training were applied to the holdout set for performance evaluation. To assess the performance of full handcrafted features, we also designed a multivariable model with Ridge L_2 regularisation on regression coefficients. Parameters were optimised using the training set and selected based on the performance in the validation set. Similar to the feature selection method, the final performance was evaluated using the holdout set.

We also assessed the complementary effect of the image score with other clinical risk factors such as biologically effective dose (BED) and histological subtypes. BED was calculated using an α/β ratio of 10 Gy, modelled as a continuous variable. We assessed the effects of two main histological subtypes, adenocarcinoma and squamous cell carcinoma and modelled them as categorical data. In the presence of the competing risk (death), Fine and Gray regression modelling was used to examine the effect of all factors on local failure. Univariate analysis was first used to confirm the significance level of each individual factor. All three variables were included in the multivariable model. For directly evaluating the effect of histological subtype between adenocarcinoma and squamous cell carcinoma variable, the model was fitted to a subset of the data

(ie, adenocarcinoma and squamous cell carcinoma patients only).

We used the multivariable regression model with Deep Profiler score and BED to both predict failure and calibrate the radiation dose to modulate the risk of local failure. *i*Gray was defined as the dose that results in a probability of failure below 5% at 24 months and is in units of BED. The calibration was achieved by estimating the cumulative incidence function (CIF) from the regression model. According to the assumptions in Fine and Gray's model, the predicted CIF can be computed for a subject with covariate vector X as follows:

$$I(t|\mathrm{X}) = 1 - [1 - I_0(t)]^{\exp(\beta^{\mathrm{T}}\mathrm{X})}$$

where $I_0(t)$ is the estimated baseline CIF,

$$\mathrm{X} = (x_{\mathrm{img}}, x_{\mathrm{BED}})^{\mathrm{T}}$$

is the covariate vector, and

$$\beta = (\beta_{\mathrm{img}}, \beta_{\mathrm{BED}})^{\mathrm{T}}$$

are the regression coefficients for image and BED covariates.

To estimate the feasibility of delivering *i*Gray-recommended doses, we permitted prescribed doses of up to 180 Gy BED for gross tumour volumes that were outside of the central zone per RTOG 0236[16] and 0618[23] (we were using the same criteria for GTVs outside the central zone as the investigators of these trials). For central tumours, we partitioned the central zone region, which is within a 2 cm radius of large airways or the proximal bronchial tree, into four equal segments. We assigned the following gradient BED schema to tumours, from the most proximal to the most distal segments: 108, 132, 149.5, and 168 Gy. 108 Gy BED (60 Gy in eight fractions) has been previously shown to be safe for ultra-central tumours.[24,25] The use of 132 Gy BED in the next segment is per RTOG 0813[26] (we were using the same BED for this segment as the investigators of this trial), which indicated that the maximal tolerated dose in patients with centrally located tumours is 12 Gy in five fractions. For 1–2 cm tumours, minimal to no overlap between the treated volume and the proximal bronchial tree and central organs at risk is expected because of more limited respiratory motion in the central zone and planning target volume expansions of only around 5 mm. Nevertheless, we used a conservative linear gradient of risk to estimate putative safe doses in these regions. These safe dose estimates are theoretical as the relationships between dose escalation, a stratified central zone, and toxicity have yet to be thoroughly investigated.

To find the voxels of an input volume that contribute the most toward the prediction of treatment failures, we took the derivative of the final partial likelihood loss with respect to the input CT volume and evaluated each volume X_i as

$$\left.\frac{\partial L_s}{\partial \mathrm{X}}\right|_{\mathrm{X}_i}$$

This derivative provides a scalar quantity for each of the voxels in the input volume, indicating the influence of the variation of voxel to the output of the model. The magnitude of these values was projected on the CT image to create a saliency map.

Statistical analysis

To quantify the predictive performance, the C-index was measured between network output and actual event (local failure) time. The C-index is a measurement between 0 and 1 that indicates how well the prediction model can order the actual event times. –1 indicates perfect concordance while 0·5 indicates no better concordance than chance. The averaged C-index across all five folds was calculated. The confidence interval was calculated using a bootstrap approach. We calculated cross-validated C-indices based on bootstrap resampling of the holdout set and repeated 1000 times. The 2·5th and 97·5th percentile of the bootstrapped C-index distribution was used as an estimation of the 95% CI.

We compared the predictive performance of handcrafted radiomics and our imaging fingerprint using the C-index. The performance of tumour two-dimensional (2D) CT size, the maximum 3D diameter, and volume were used as comparators. We applied a bootstrap method to compare different models. For each model, we randomly resampled the holdout set and calculated the C-index. This was repeated 100 times for all five folds. The Wilcoxon signed rank test was used to compare the C-index distributions of different models.

To further explore the association between the imaging index and failure time, we did a competing risk analysis to estimate the cumulative incidence of local failure. The Kaplan-Meier method is not appropriate for estimating the incidence of therapy failure in the presence of death, because patient death leads to the censoring of the primary outcome. As mortality is not completely independent from therapy failure, death without evidence of local failure was treated as a competing event. The median score in the training set was computed and then applied as a threshold to stratify patients in the holdout set into high-risk and low-risk groups. After the cross-validation was complete, each patient was classified into one of the risk groups. Cumulative incidence curves were estimated for each group, and Gray's test was used to determine the significance of difference between two curves.[27] Statistical analyses were done with R (version 3.2.5).[28]

Role of the funding source

This work was supported, in part, by Siemens Medical Solutions USA. Siemens developed Deep Profiler and

	Internal study cohort (n=849)	Independent validation cohort (n=95)	p value
Follow-up (months)	20·9 (11·0–38·0)	16·4 (11·4–24·6)	0·0016
Age (years)	74·1 (67·6–80·7)	76·0 (70·0–82·3)	0·044
Sex	..	..	0·16
Female	440 (51%)	57 (60%)	..
Male	409 (48%)	38 (40%)	..
Treated tumour size (cm)	2·3 (1·6–3·4)	1·8 (1·3–2·7)	0·00013
Overall tumour stage			
Number of tumours	849	102	0·29
I	645 (76%)	78 (76%)	..
II	81 (10%)	10 (10%)	..
III	8 (1%)	0	..
IV	74 (9%)	5 (5%)	..
Recurrent	41 (5%)	9 (9%)	..
Histology	..	..	0·21
Number of tumours	849	102	
Adenocarcinoma	255 (30%)	43 (42%)	..
Squamous cell carcinoma	248 (29%)	24 (24%)	..
NSCLC-NOS	47 (6%)	8 (8%)	..
Neuroendocrine carcinoma	14 (2%)	1 (1%)	..
Other	22 (3%)	1 (1%)	..
Non-diagnostic biopsy	74 (9%)	6 (6%)	..
No biopsy	189 (22%)	19 (19%)	..
Indications for treatment	..	..	0·52
Number of tumours	849	102	
Definitive	738 (87%)	88 (86%)	..
Salvage	52 (6%)	9 (9%)	..
Oligometastatic	50 (5%)	5 (5%)	..
Other	9 (1%)	0	..
Total radiation dose (Gy)	50 (30–60)	50 (34–60)	0·017
Fractions	5 (1–10)	5 (1–5)	0·075
Biologically effective dose (Gy)	100 (39–180)	100 (72–180)	0·0071

Data are median (IQR) or n (%) with the exception of total radiation dose, fractions, and biologically effective dose, which are represented by median (range). NSCLC-NOS=non-small-cell lung cancer-not otherwise specified. p values were calculated with Pearson's χ^2 test for categorical variables and with the non-parametric test of medians for continuous variables. A p value below 0·05 was considered significant.

Table 1: **Baseline characteristics of the study population**

contributed to data analysis, data interpretation, and the writing of aspects of the manuscript. Siemens had no role in data collection or the overall experimental design of the study. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

A total of 1275 patients were assessed for eligibility and 944 met our eligibility criteria; 849 were in the internal study cohort and 95 were in the independent validation cohort (table 1). There were 89 local failures overall in the internal study cohort. The number of local failures was on average 56 in the training set, 15 in the validation set, and 18 in the holdout set. There were 16 local failures in the independent validation cohort. The cumulative incidence of local failure at 3 years was 13·5% (95% CI 10·8–16·2) in the internal study population. Patients in the internal study cohort were stratified into high-risk and low-risk groups on the basis of a median score cut-off of a neural network-derived imaging signature from the training set. This process was repeated for each partition and the results of all five folds were concatenated for statistical analysis. 469 (55%) of 849 patients were stratified into the high-risk group and 380 (45%) patients were stratified into the low-risk group. Estimated cumulative incidence curves of local failure for the internal study cohort and each risk group are shown in figure 2. Gray's test for equality across Deep Profiler risk groups was significant (p<0·0001). Radiotherapy was more successful in patients in the low-risk group than in patients in the high-risk group; 3-year cumulative incidence of local failure was 5·7% (95% CI 3·5–8·8) in the low-risk group compared with 20·3% (16·0–24·9) in the high-risk group (hazard ratio [HR]=3·64 [95% CI 2·19–6·05], p<0·0001).

To establish the clinical setting in which Deep Profiler can be most predictive of local failure, we examined the effects of tumour stage on Deep Profiler's prediction accuracy. Scores varied based on tumour stage, with stage IA tumours having the lowest mean score (figure 2). Despite differences in the mean scores of Deep Profiler across some stages of disease, there was significant variation within and across individual stages. These results suggested that Deep Profiler had learned information beyond tumour stage. Consistent with this observation, Deep Profiler predicted local failure in patients with early-stage or late-stage cancers (figure 2).

To assess the effect of possible variation in the types of treatments delivered or CT image acquisition, we assessed the effect of motion management, use of adjuvant chemotherapy, and CT scanner type on Deep Profiler. In 96 (11·3%) of 849 cases, motion could not be adequately restricted to less than 1 cm. This necessitated an Active Breathing Coordinator. Scores were not significantly different based on the type of motion management used for treatment (Active Breathing Coordinator *vs* abdominal compression, p=0·353; appendix). Scores were, however, significantly higher in patients who received adjuvant chemotherapy (8·5% of patients received adjuvant chemotherapy; p<0·0001), although there was significant overlap across the two groups (appendix). These data are consistent with the physician-directed recommendation of adjuvant chemotherapy on the basis of variables that are perceived to lead to a higher risk of treatment failure[29] and suggest that Deep Profiler could potentially inform these recommendations. Four distinct CT scanners were used for the internal study cohort (see Methods) and the number of patients scanned on each of the scanners was 499 for CT-1, 244 for CT-2, 40 for CT-3, and 61 for CT-4. The identity of the scanner could not be definitively determined for five patients. Lastly, scores obtained from the two most frequently used CT scanners (both of which are Philips Big

Bore) had similar accuracy in predicting local failures (appendix).

We compared the performance of our neural network-derived imaging index to 2D CT size, maximum 3D CT size, 3D tumour volume, and classical radiomic features. Our learning-based framework was superior to classical radiomics features, which were in turn superior to tumour volume followed by 2D size values (table 2). The superiority of Deep Profiler indicates that features beyond tumour size, which have previously been associated with local failure after high-dose radiotherapy to the lung,[30] can be identified using our deep learning algorithm.

On univariate analysis, a higher image-based risk score (Deep Profiler), lower radiation dose, and histological subtype were associated with an increased risk of local failure (appendix). On multivariable analyses, all three factors remained significantly associated with local failure (table 3). The multivariable models that included Deep Profiler and clinical variables predicted treatment failures with a C-index of 0·72 (95% CI 0·67–0·77), which was a significant improvement when compared with classical radiomics (p<0·0001) or 3D volume (p<0·0001). These results indicated that an image-based score can provide complementary information to the clinically established variables of histological subtype and radiation dose.[19]

We hypothesised that treatment failures can be mitigated by higher radiation doses and that we can model this relationship to guide dose individualisation. First, we built a Fine and Gray's regression model using the imaging signature and dose of radiation. This enabled us to model the risk of local recurrence by tuning the dose of radiation accordingly. Importantly, the type of treatment delivered is the only mutable variable identified in the model; tumour size, CT image features, and histology are fixed. Using this model, we calculated the probability of local failure at 24 months after treatment as a function of radiation dose. Our results indicated that local failure can be significantly reduced as a function of radiation dose (figure 3).

We then calculated *i*Gray, the patient-specific dose that reduces the probability of treatment failure to below 5%, for each patient using a permuted holdout set design.

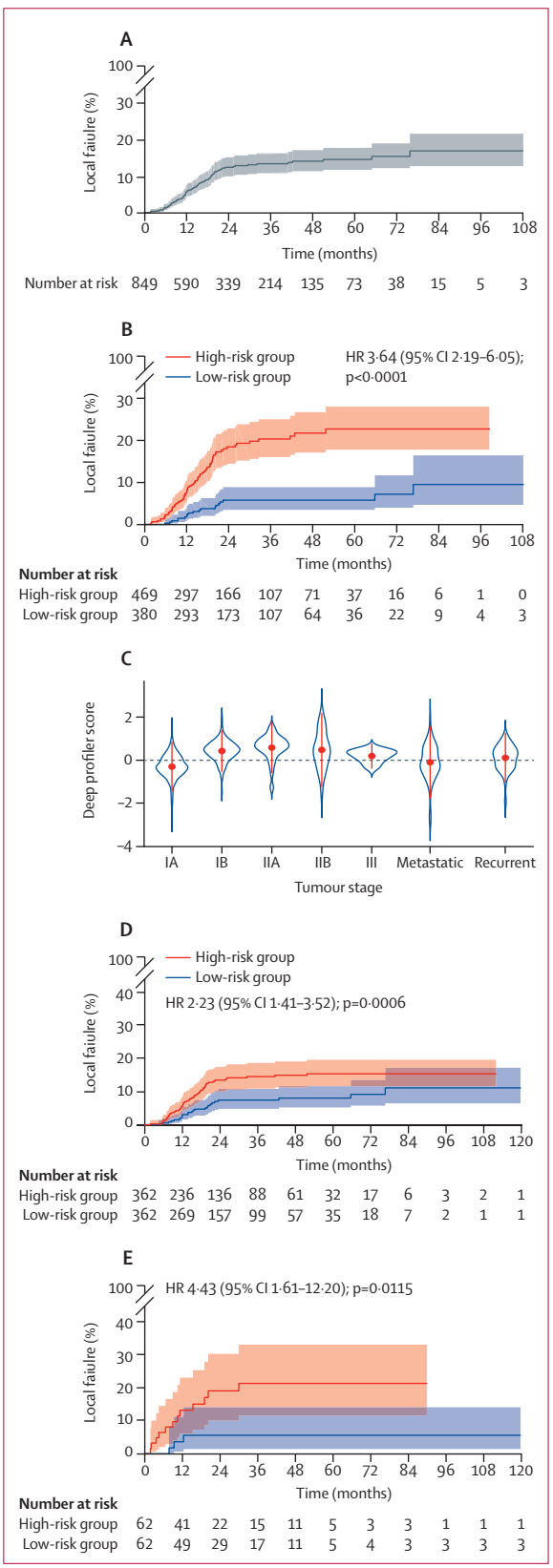

Figure 2: **Deep Profiler predicts local failure in the internal study cohort**
(A) Estimated cumulative incidence curves of local failure. The shaded area represents 95% CIs. (B) Estimated cumulative incidence curves of local failure, stratified into high-risk and low-risk groups on the basis of the median Deep Profiler score obtained with each cross-validation step followed by concatenation. The shaded area represents 95% CIs. (C) Violin plot of Deep Profiler scores by tumour overall stage. The mean (point) and SDs (point range) are shown. (D) Estimated cumulative incidence curves of local failure apportioned by the median Deep Profiler score for early-stage (IA–IIB) tumours. The shaded area represents 95% CIs. (E) Estimated cumulative incidence curves of local failure apportioned by the median Deep Profiler score for late-stage (III, IV, and recurrent) tumours. The shaded area represents 95% CIs. HR=hazard ratio.

	Concordance index (95% CI)	p value
Deep Profiler	0·71 (0·67–0·77)	1 (ref)
Two-dimensional CT size	0·61 (0·55–0·67)	$2·05\times10^{-32}$
Maximum 3D diameter	0·66 (0·60–0·71)	$1·06\times10^{-20}$
Three-dimensional volume	0·67 (0·61–0·73)	$9·78\times10^{-14}$
Classical radiomics (feature selection)	0·65 (0·60–0·71)	$4·23\times10^{-25}$
Classical radiomics (regularisation)	0·68 (0·63–0·74)	$1·18\times10^{-10}$

Table 2: **Prognostic performance of different models of outcome prediction**

	Hazard ratio (95% CI)	p value
Deep Profiler signature	1·65 (1·02–2·66)	0·042
BED (continuous)	0·98 (0·97–0·99)	0·026
Adenocarcinoma versus Squamous cell carcinoma	0·494 (0·28–0·87)	0·029
Adenocarcinoma versus others	0·52 (0·29–0·93)	0·027

Table 3: **Predictors of local failure in multivariable analysis**

The kernel densities of dose delivered compared with *i*Gray showed significant overlap (figure 3). The ranges of *i*Gray and dose delivered were 21·15–277·1 Gy and 39–180 Gy. The variance of the distributions of *i*Gray and dose delivered were 40·6 and 30 Gy. The percent dose difference required to achieve *i*Gray for each patient was calculated and their distributions were plotted for a function of each dose delivery interval (figure 3). A dose reduction was recommended in 23·3% of patients (*i*Gray < actual dose delivered). Together, these results indicated that *i*Gray doses had wider range, greater variance, were more continuous than actual doses delivered (curve in figure 3C), and recommended dose de-escalation in a subset of patients.

To assess the feasibility of delivering *i*Gray dose recommendations, we first estimated the effect of incremental dose increases on the probability of local failure in patients who received a BED of 100 Gy, the most frequent treatment dose in the internal study cohort (n=445). We used our model to estimate local failure probabilities at 24 months (figure 3) and showed that even incremental increases in the dose delivered to these patients can reduce treatment failure probability. To generate an estimate of the extent of feasibility in all patients in our internal study cohort, we used a gradient dose scheme that was extrapolated from previous dose escalation studies to avoid airway toxicity based on the proximity of the tumour to the proximal bronchial tree. Using this scheme, the cumulative relative frequency of safely achieving *i*Gray was 63·5% (figure 3).

To examine the agreement between the observed outcomes and the multivariable model with *i*Gray and BED, we calculated calibration curves. A calibration curve was obtained by plotting the average predicted probability at 1, 2, or 3 years after radiation treatment against cumulative incidence curve estimates of the actual outcome (figure 4). The calibration curve showed that our model accurately predicts treatment outcomes.

To measure the effect of dataset size on prediction accuracy, we randomly selected 60% of the patients in the dataset and calculated the concordance indices using our deep learning platform and a classical clinical risk factor (ie, volume), then compared our results to analyses using 100% of the patients. We found that although volume appears to reach a plateau in accuracy, our framework's performance is significantly higher with increases in sample size (figure 4). These results establish the scalability of deep learning in our dataset and suggest that improvements in accuracy are more likely with deep learning-based approaches than with tumour volume measurements in terms of dataset growth.

To assess the effect of each voxel on treatment failure, we calculated a saliency map for each tumour. Saliency projects a weight in heatmap form to each voxel in the image and this weight reflects the importance of that voxel in the image risk score (figure 5). Crucially, the most salient voxels were within the gross tumour volumes and planning target volumes across all tumours, indicating that the gross tumour and peritumoral regions are the most relevant voxels to the model (figure 5). This saliency detection method is also a crucial examination of the spatial sampling methodology of cropping to a 64×64×32 subvolume encompassing the tumour, indicating that the volume is sufficiently large to encompass the most salient voxels but not too broad to result in a classifier that fails to understand despite having high accuracy (eg, spurious voxel associations).[31] Lastly, there were a number of salient voxels outside of the gross tumour volumes (37·8%) and planning target volumes (20·5%) across the dataset. The role of these outlying salient voxels in marginal treatment failures remains unclear.

We sought to measure the accuracy of Deep Profiler using a different but plausibly related independent cohort of patients who received stereotactic body radiotherapy to the lung. A total of 95 patients with 102 tumours (metachronous or synchronous treatments were included) from seven affiliate treatment centres met our eligibility criteria. The number of patients scanned in the independent validation cohort were 14 with GE Medical Systems Discovery ST, 22 with Philips Brilliance CT Big Bore, 17 with Philips Gemini GXL, and 49 with a Siemens SOMATOM Definition AS, indicating substantial diversity in the type of CT scanner used. Differences in baseline patient characteristics compared with our internal study cohort included a shorter median time to follow-up, smaller tumours, lower radiation doses, and older median age (table 1). These results showed differences between the internal and independent validation cohorts, hence allowing for an assessment of both the reproducibility and transportability of the model.[32]

There were 16 local failures in the independent validation cohort. The cumulative incidence of local failure in this population at 2 years was 19·7% (95% CI 10·9–30·4). Patients were stratified into high-risk and

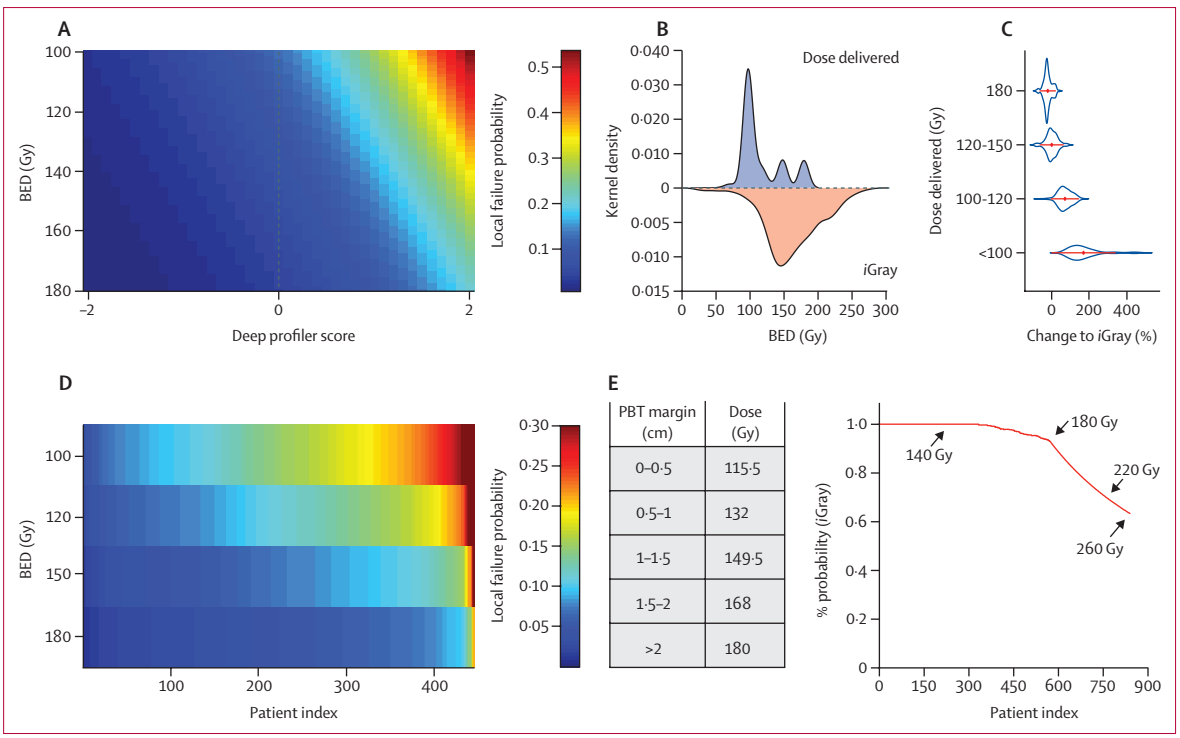

Figure 3: **Ameliorating treatment failure by adjusting the radiation dose**
(A) Heatmap of the relationship between Deep Profiler score (*x*-axis), *i*Gray in BED units (*y*-axis), and the probability of local failure (right legend). As *i*Gray BED increases, the risk of failure decreases and vice versa. (B) Kernel density estimation of the actual dose of radiation delivered and *i*Gray. (C) Violin plots of the distribution for the percent change in dose delivered required to achieve iGray stratified by the four treatment categories (*y*-axis). (D) Heatmap of the relationship between patient index (*x*-axis), *i*Gray in BED units (*y*-axis), and the probability of local failure (right legend). The probability of local failure was calculated for patients receiving the most common treatment regimen of 100 Gy BED after incremental increases in radiation dose. The overall rate of local failure is significantly diminished at the highest doses. (E) A proposed theoretical dose scheme to avoid proximal airway toxicity, and the cumulative relative frequency of feasibly delivering *i*Gray according to this proximal dose scheme. Some *i*Gray doses are shown, to highlight the range in which a substantial decrement in feasibility occurs. BED=biologically effective dose. PBT=proximal bronchial tree.

low-risk groups on the basis of a median Deep Profiler score from a training dataset that included all 849 patients in our internal study cohort. Estimated cumulative incidence curves of local failure for each risk group are shown in figure 6. Gray's test for equality across Deep Profiler risk groups was significant (p=0·002). Radiotherapy was more successful in patients in the low-risk group than in patients in the high-risk group; 2-year cumulative incidence of local failure was 9·5% (95% CI 2·8–21·3) in the low-risk group and 39% (19·6–58·1) in the high-risk group. Deep Profiler predicted treatment failure with a C-index of 0·77 (95% CI 0·66–0·92), which was calculated on the basis of bootstrap resampling of the external dataset and repeated 1000 times. These results show that Deep Profiler can predict treatment failures accurately across diverse clinical settings and distinct CT simulator scanners (table 1).

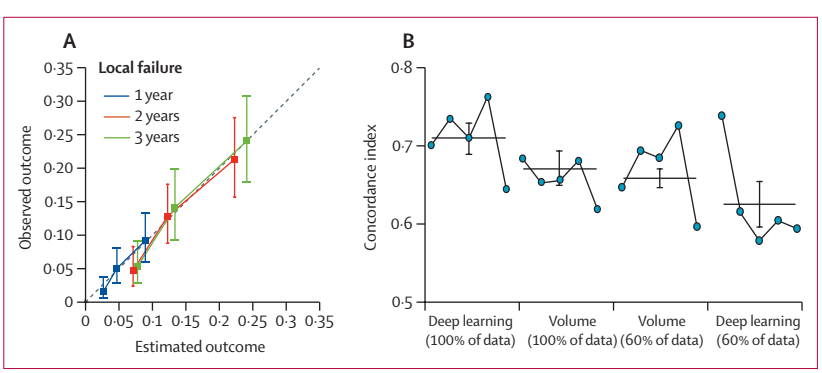

Figure 4: **Model performance and the scalability of the deep neural network**
(A) Calibration curves to examine the agreement between the observed outcomes and the multivariable model with *i*Gray and BED. A calibration curve was obtained by plotting the average predicted probability at 1, 2, or 3 years after radiation treatment against cumulative incidence curve estimates of the actual outcome. Vertical bars represent 95% CIs for observed local failure probabilities. (B) Comparison of model performance using 60% of the dataset and 100% of the dataset shows the superiority and the potential scalability of the deep learning algorithm. Horizontal lines and error bars represent the mean concordance indices and SEs across five folds.

Discussion

We developed a new deep learning model that predicts local tumour failure probability using pretreatment CT images of patients treated with stereotactic body radiation therapy for lung tumours and an integrated imaging and clinical predictor that estimates the biologically effective dose of radiation required to achieve tumour local control. Accurate estimates of the probability of treatment success for individual patients can substantially improve clinical

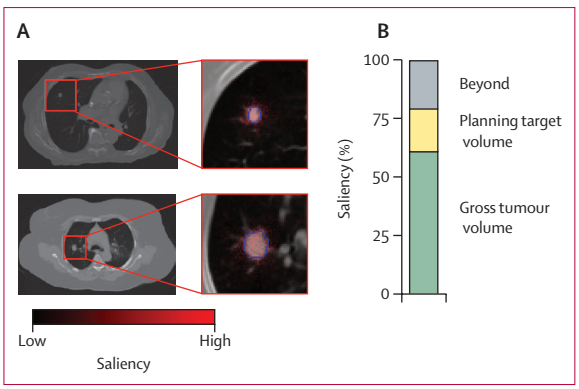

Figure 5: **Tumour saliency extends beyond the physician delimited tumour volume**
(A) A saliency map represents an overlay of heatmap colour (red=high saliency, black=low saliency) on voxels in the CT image region of interest. Blue contour lines represent the delineation of tumour volume. Two patient examples are shown. (B) The proportion of voxels with saliency greater than 0·20 within the gross tumour volume, the planning target volume, or beyond both volumes are shown.

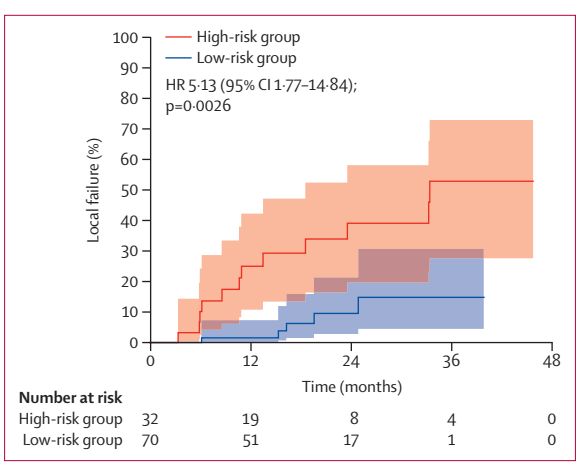

Figure 6: **Deep Profiler predicts local failure in the independent validation cohort**
Estimated cumulative incidence curves of local failure apportioned by the median Deep Profiler score derived from the internal study cohort. The shaded area represents 95% CIs. Patients were stratified into high-risk and low-risk groups on the basis of the median Deep Profiler score obtained with each cross-validation step followed by concatenation. HR=hazard ratio.

outcomes. In this study, we show that the clinical responses of cancer to radiotherapy vary in a manner not fully explained by clinical and histopathological variables alone, and that CT image-based features contribute to this variance. The most important message in our study is that predictive features can be learned from CT images and can contribute to the individualisation of radiation dose.

To date, quantitative image analyses have not been used to personalise delivery of cancer treatment.[33–35] To address these limitations, we trained a neural network to learn the multidimensional feature space of CT images taken from a large cohort of patients with cancers in the lung who were receiving treatment with a range of radiation doses. Unlike unconstrained machine learning approaches to imaging data, we used a multitask approach that takes advantage of established image-based radiomic features to partially delimit and inform the neural network. We showed that the Deep Profiler approach, is superior to deep learning or classical radiomics alone at accurately predicting radiotherapy treatment failures in varied clinical settings.

To limit spurious voxel associations with our predicted outcome, we incorporated previous knowledge by using the physician-delimited tumour volume as part of the input into our network. Although manual annotation could bias the feature extractions, we showed that the voxels that are most deterministic for treatment failure localise within the physician-contoured volumes (gross tumour volumes or planning target volumes). Conversely, some salient voxels localised to the peritumoral regions or tumour margin. Since classical radiomics approaches disregard image information outside of the gross tumour volume, the identification of these salient voxels is an additional advantage of this approach. The potential association of these voxels with marginal recurrences remains to be explored. To the extent that marginal salient voxels are predictive of local failures, automatic contouring of tumour saliency maps could represent a leap toward more accurate tumour volume delineation and informed personalised dose delivery.[36,37]

Our results have several additional clinical implications. Firstly, we are able to find image-distinct subpopulations with differential sensitivity to radiotherapy; Deep Profiler scores are significantly associated with treatment failures across varied clinical settings, including distinct stages of disease, CT simulation scanners, and longitudinal periods. Secondly, we provide an integrated method that combines images and established clinical variables to individualise radiation dose. *i*Gray uses the clinically validated linear quadratic model,[38] is empirically derived in that no assumptions are made regarding individual tumour radiosensitivity (α and β in the tumour toxicity isoeffect remain constant), and its output is directly clinically actionable by recommending a dose that can be achieved using several treatment schedules. Finally, our prediction accuracy could evolve. A crucial feature of neural network-based prediction is the potential for substantial improvements in accuracy with scale, in contrast with other computational methods such as classical radiomics, whose accuracy appears to plateau in the early phases of dataset growth. As our dataset increases in size and is augmented by integration into large data-sharing collaborations, the network is expected to substantially improve in prediction accuracy.

Another important element of the evolvability of our model is the eventual stratification of the dataset into more homogeneous populations, on the basis of variables such as cancer subtype, clinical stage, and use of systemic adjuvant therapy. This is particularly compelling considering the preliminary efficacy of stereotactic body radiotherapy in patients with oligometastatic disease

from distinct cancer types.[39,40] The use of stereotactic body radiotherapy in varied clinical settings will result in larger and more diverse datasets that are more amenable to data partitions, and therefore improved model accuracy.

An image-based framework for the personalisation of radiotherapy dose can substantially alter the clinical radiotherapy paradigm. Radiation oncologists can calibrate the dose of radiation delivered on the basis of the risk of treatment failure, which itself is a continuum. This is an advantage which largely mitigates binary decisions about treatment and instead permits the adjustment of radiation treatments to prevent treatment failures. *i*Gray can assist in the design of image-stratified, radiotherapy-based trials; in this role, it can guide the evolution of radiotherapy towards dose-delivery strategies that are calibrated on the basis of individual predictions of tumour control probability.

There are several characteristics of Deep Profiler and *i*Gray that suggest minimal implementation barriers. Because of the strict requirement for the acquisition of radiation planning CT images for radiotherapy, each radiation treatment centre is likely to have an extensive CT dataset that could be used for model development and implementation. Combined with the automated feature algorithms of scalable deep learning-based prediction platforms, this represents an opportunity to directly improve medical-decision support for diverse patient populations receiving radiotherapy.

The strengths of our study include the large number of patients evaluated, the completeness of the dataset, the use of a carefully annotated, radiotherapy-specific outcome (local failure) rather than a surrogate of treatment failure (eg, progression-free survival, cancer-specific mortality, or overall survival), and the use of a readily implementable and highly tractable image-based score as a backbone for our analyses. The limitations of our study include our inability to fully account for all potential causes of bias, the explicit population heterogeneity in our datasets (eg, clinical stage, radiation dose, CT scanners, motion management), and the limited size of the independent validation cohort. We also did not account for normal tissue toxicity. These limitations can be addressed in part with the addition of new datasets and emerging tools that predict lung toxicity.[41]

In summary, we combined clinical variables with deformable radiomic features through the deep learning of CT imaging-based features, to create a clinically meaningful unit called *i*Gray which allows for individualised radiation dose delivery. This framework could be readily implemented for pretreatment risk stratification and risk-adapted dose optimisation in clinical trials and, ultimately, in routine clinical practice.

Contributors

DW, MG, TZ, and SD assisted with data generation and analysis. MG and SD assisted with the writing of the research in context section and edited the manuscript. BL did the training and optimisation of the neural network, analysed and interpreted the data, and contributed to the writing of the methods and results. NM and LL assisted with interpretation and edited the manuscript. AK supervised the training and optimisation of the neural network, assisted with interpretation, and edited the manuscript. MEA conceived, designed, analysed, interpreted, and supervised the overall work and wrote the manuscript.

Declaration of interests

BL, AK, NM, LL, and MEA are named inventors in a patent pending for the use of Deep Profiler and *i*Gray to personalise radiotherapy dose. MEA receives grant support, travel support, and honoraria from Bayer AG, and receives grant support from Siemens Medical Solutions USA. MEA is also supported by the National Institutes of Health (grant numbers KL2 TR0002547 and R37 CA222294), the American Lung Association, and VeloSano. BL, LL, NM, and AK report personal fees from Siemens Medical Solutions USA. NM reports personal fees from Siemens Healthcare. All other authors declare no competing interests.

Data sharing

The datasets analysed in this study will be available from the corresponding author (abazeem@ccf.org) at the time of publication. Per institutional policy, the datasets are designated limited access. Upon receiving access, the investigator may only use them for the purposes outlined in the request to the data provider, and redistribution of the data is prohibited. Deep Profiler and *i*Gray are based on research and are not currently commercially available. Due to regulatory reasons, their future availability cannot be guaranteed.

Acknowledgments

This work was supported, in part, by Siemens Medical Solutions USA.

References

1 Morin O, Vallieres M, Jochems A, et al. A deep look into the future of quantitative imaging in oncology: a statement of working principles and proposal for change. *Int J Radiat Oncol Biol Phys* 2018; **102:** 1074–82.
2 National Research Council. Mathematics and physics of emerging biomedical imaging. Washington, DC: National Academy Press, 1996.
3 Barten PGJ. Physical model for the contrast sensitivity of the human eye. *Proc SPIE* 1992; **1666** (abstract). https://doi.org/10.1117/12.135956.
4 Barten PGJ. Contrast sensitivity of the human eye and its effects on image quality. PhD thesis, Technische Universiteit Eindhoven, 1999.
5 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42:** 60–88.
6 Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017; **14:** 749–62.
7 Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012; **48:** 441–46.
8 Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Communications* 2014; **5:** 4006.
9 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.
10 Samei E, Rowberg A, Avraham E, Cornelius C. Toward clinically relevant standardization of image quality. *J Digit Imaging* 2004; **17:** 271–78.
11 Budin-Ljøsne I, Burton P, Isaeva J, et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. *Public Health Genomics* 2015; **18:** 87–96.
12 Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013; **45:** 1113–20.
13 Yard BD, Adams DJ, Chie EK, et al. A genetic basis for the variation in the vulnerability of cancer to DNA damage. *Nat Commun* 2016; **7:** 11428.
14 Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24:** 1559–67.
15 Causey JL, Zhang JY, Ma SQ, et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci Rep* 2018; **8:** 9286.
16 Timmerman R, Paulus R, Galvin J, et al. Stereotactic body radiation therapy for inoperable early stage lung cancer. *JAMA* 2010; **303:** 1070–76.

17 Timmerman RD, Hu C, Michalski J, et al. Long-term results of RTOG 0236: a phase II trial of stereotactic body radiation therapy (SBRT) in the treatment of patients with medically inoperable stage I non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* 2014; **90:** S30.

18 Videtic GM, Hu C, Singh AK, et al. A randomized phase 2 study comparing 2 stereotactic body radiation therapy schedules for medically inoperable patients with stage I peripheral non-small cell lung cancer: NRG Oncology RTOG 0915 (NCCTG N0927). *Int J Radiat Oncol Biol Phys* 2015; **93:** 757–64.

19 Woody NM, Stephans KL, Andrews M, et al. A histologic basis for the efficacy of SBRT to the lung. *J Thorac Oncol* 2017; **12:** 510–19.

20 Hörner-Rieber J, Bernhardt D, Dern J, et al. Histology of non-small cell lung cancer predicts the response to stereotactic body radiotherapy. *Radiother Oncol* 2017; **125:** 317–24.

21 Baine MJ, Verma V, Schonewolf CA, Lin C, Simone CB 2nd. Histology significantly affects recurrence and survival following SBRT for early stage non-small cell lung cancer. *Lung Cancer* 2018; **118:** 20–26.

22 van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017; **77:** e104–07.

23 Timmerman RD, Paulus R, Pass HI, et al. Stereotactic body radiation therapy for operable early-stage lung cancer: findings from the NRG Oncology RTOG 0618 trial. *JAMA Oncol* 2018; **4:** 1263–66.

24 Murrell DH, Laba JM, Erickson A, Millman B, Palma DA, Louie AV. Stereotactic ablative radiotherapy for ultra-central lung tumors: prioritize target coverage or organs at risk? *Radiat Oncol* 2018; **13:** 57.

25 Kimura T, Nagata Y, Harada H, et al. Phase I study of stereotactic body radiation therapy for centrally located stage IA non-small cell lung cancer (JROSG10-1). *Int J Clin Oncol* 2017; **22:** 849–56.

26 Bezjak A, Paulus R, Gaspar LE, et al. Safety and efficacy of a five-fraction stereotactic body radiotherapy schedule for centrally located non-small-cell lung cancer: NRG Oncology/RTOG 0813 trial. *J Clin Oncol* 2019; **37:** 1316–25.

27 Scrucca L, Santucci A, Aversa F. Regression modeling of competing risk using R: an in depth guide for clinicians. *Bone Marrow Transplant* 2010; **45:** 1388–95.

28 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2018.

29 Ernani V, Appiah AK, Marr A, et al. Adjuvant systemic therapy in patients with early-stage NSCLC treated with stereotactic body radiation therapy. *J Thorac Oncol* 2019; **14:** 475–81.

30 Allibhai Z, Taremi M, Bezjak A, et al. The impact of tumor size on outcomes after stereotactic body radiation therapy for medically inoperable early-stage non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* 2013; **87:** 1064–70.

31 Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA; Aug 13–17, 2016. 1135–44. https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf (accessed June 10, 2019).

32 Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; **68:** 279–89.

33 Huynh E, Coroller TP, Narayan V, et al. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiother Oncol* 2016; **120:** 258–66.

34 Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* 2015; **114:** 345–50.

35 Parmar C, Leijenaar RT, Grossmann P, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci Rep* 2015; **5:** 11044.

36 Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS Med* 2018; **15:** e1002711.

37 Huang ZL, Wang XG, Wang JS, Liu WY, Wang JD. Weakly-supervised semantic segmentation network with deep seeded region growing. *Proc Cvpr Ieee* 2018: 7014–23. http://openaccess.thecvf.com/content_cvpr_2018/html/Huang_Weakly-Supervised_Semantic_Segmentation_CVPR_2018_paper.html (accessed June10, 2019.)

38 Brenner DJ. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Semin Radiat Oncol* 2008; **18:** 234–39.

39 Palma DA, Olson R, Harrow S, et al. Stereotactic ablative radiotherapy versus standard of care palliative treatment in patients with oligometastatic cancers (SABR-COMET): a randomised, phase 2, open-label trial. *Lancet* 2019; **393:** 2051–58.

40 Gomez DR, Blumenschein GR Jr, Lee JJ, et al. Local consolidative therapy versus maintenance therapy or observation for patients with oligometastatic non-small-cell lung cancer without progression after first-line systemic therapy: a multicentre, randomised, controlled, phase 2 study. *Lancet Oncol* 2016; **17:** 1672–82.

41 Cunliffe A, Armato SG 3rd, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *Int J Radiat Oncol Biol Phys* 2015; **91:** 1048–56.