Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning

Zheng, Sunyi; Guo, Jiapan; Langendijk, Johannes A; Both, Stefan; N J Veldhuis, Raymond; Oudkerk, Matthijs; van Ooijen, Peter M A; Wijsman, Robin; Sijtsema, Nanna M

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2023.109483

Publication date:
2023

[Article Full Title]
Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning

[Author Names]
Sunyi Zheng, PhD [a, b, c], Jiapan Guo, PhD [a, d], Johannes A. Langendijk, MD, PhD [a], Stefan Both, PhD [a], Raymond N. J. Veldhuis, PhD [e], Matthijs Oudkerk, MD, PhD [f], Peter M.A. van Ooijen, PhD [a], Robin Wijsman, MD, PhD [a], Nanna M. Sijtsema, PhD [a]*

[Author Institutions]
a Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, The Netherlands; b Artificial Intelligence and Biomedical Image Analysis Lab, School of Engineering, Westlake University; c Institute of Advanced Technology, Westlake Institute for Advanced Study; d Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen; e Faculty of Electrical Engineering, University of Twente, The Netherlands; f Faculty of Medical Science, University of Groningen, The Netherlands

[Corresponding Author Name & Email Address]
Sunyi Zheng, PhD; E-mail: zhengsunyi@westlake.edu.cn; Postal address: Shilongshan Road No.18, Cloud Town, Xihu District, Hangzhou, Zhejiang, China.

Short title: OS prediction for stage I-IIIA NSCLC using DL

Abstract:

Background and purpose: The aim of this study was to develop and evaluate a prediction model for 2-year overall survival (OS) in stage I-IIIA non-small cell lung cancer (NSCLC) patients who received definitive radiotherapy by considering clinical variables and image features from pre-treatment CT-scans.

Materials and methods: NSCLC patients who received stereotactic radiotherapy were prospectively collected at the UMCG and split into a training and a hold out test set including 189 and 81 patients, respectively. External validation was performed on 228 NSCLC patients

who were treated with radiation or concurrent chemoradiation at the Maastro clinic (Lung1 dataset). A hybrid model that integrated both image and clinical features was implemented using deep learning. Image features were learned from cubic patches containing lung tumours extracted from pre-treatment CT scans. Relevant clinical variables were selected by univariable and multivariable analyses.

Results: Multivariable analysis showed that age and clinical stage were significant prognostic clinical factors for 2-year OS. Using these two clinical variables in combination with image features from pre-treatment CT scans, the hybrid model achieved a median AUC of 0.76 [95% CI: 0.65-0.86] and 0.64 [95% CI: 0.58-0.70] on the complete UMCG and Maastro test sets, respectively. The Kaplan-Meier survival curves showed significant separation between low and high mortality risk groups on these two test sets (log-rank test: p-value < 0.001, p-value = 0.012, respectively)

Conclusion: We demonstrated that a hybrid model could achieve reasonable performance by utilizing both clinical and image features for 2-year OS prediction. Such a model has the potential to identify patients with high mortality risk and guide clinical decision making.

Running title: OS prediction for stage I-IIIA NSCLC using DL

Keywords: lung cancer; Radiotherapy; deep learning; logistic regression; overall survival

Introduction

Lung cancer is one of the deadliest cancer types in the developing and developed countries. The 5-year survival rate of non-small cell lung cancer (NSCLC) is only around 20% [1]. Radiotherapy combined with chemotherapy or immunotherapy have increased the overall survival (OS) rate of NSCLC patients compared to the use of chemotherapy or immunotherapy alone [2]. However, treatment approaches for NSCLC largely depend on tumour stage and

treatment outcomes vary widely among individual patients, e.g. the 2-year overall survival rate after Stereotactic Body Radiation Therapy ranges from 50% to 71% [3]. Therefore, an accurate prediction of overall survival is of importance to support decision-making for the most optimal treatment options.

To guide clinical decision making, clinical models such as nomograms have been developed to estimate individual risks for patients with NSCLC using clinical factors. Louie et al. [4], developed a nomogram for 5-year OS in NSCLC patients including age, World Health Organization performance status, smoking status, tumour size and Charlson Comorbidity index (CCI). The proposed nomogram showed reasonable performance with a c-index of 0.66 on the test set. Kang et al. [5] developed prognostic nomograms for 5-year OS based on several clinical parameters of which tumour size, immune-inflammation index, diffusing capacity of carbon monoxide and CCI were the most important prognostic factors. The c-index of the developed nomograms was 0.72 on the hold-out set from the same centre. However, no external validation was performed.

Recently, advanced deep learning techniques have shown their effectiveness in automated quantitative image analysis for lung cancer, including early detection [6] and cancer risk assessment [7]. The use of deep learning also promotes the development of automatic OS prediction [8-10]. For example, to select NSCLC patients with high mortality risks, a shallow convolutional neural network was implemented [9]. Trained on the data collected from multiple centres, the proposed network showed a c-index of 0.62 and 0.67 on the independent Maastro test set for stage I-IV and I-II lung tumours, respectively. Moreover, Hosny et al. [10] reported an AUC of 0.70 of a 3D convolutional neural network using pre-treatment CT scans from two centres for training and fine-tuning in stage I-IIIB NSCLC cancer patients treated with

radiotherapy for 2-year OS. This network was able to stratify patients with stage III NSCLC into low and high risk for mortality [8]. To date, for mortality risk stratification in stage I-IIIA NSCLC patients treated with radiotherapy, while deep learning studies mainly focused on prognostic image features, the added value of clinical parameters were not considered in the deep learning models.

Therefore, the purpose of this study was to develop and validate a hybrid deep learning-based model that considers both clinical factors and image features from pre-treatment CT-scans for the prediction of 2-year OS of stage I-IIIA NSCLC patients treated with definitive radiotherapy. Such a model could be used to stratify patients into low and high mortality risk groups.

Materials and methods

Study cohorts

The study performed a retrospective analysis of prospectively acquired NSCLC patient data available at University Medical Center Groningen (UMCG). Ethical approval was waived by the medical ethical committee of UMCG in view of the retrospective nature of the study. We used two independent radiotherapy datasets, one from UMCG and one from Maastro [11], which comprised a total of 498 patients with primary NSCLC from stage I to IIIA for this analysis. The UMCG cohort consisted of 270 patients treated in the UMCG between October 2013 and September 2018 with Stereotactic Body Radiation Therapy (SBRT). Patients with a history of other cancers were excluded, leaving only patients with (suspicion of) NSCLC. The UMCG cohort was randomly split into a training set and a test set (70% vs 30%), of which the UMCG test set was a hold-out test set used for validation. The endpoint of this study was the prediction of 2-year OS after the time of treatment initiation. OS was dichotomised at the 2-

year mark for classification and patients that were censored before the 2-year mark were excluded. The Maastro test set [11] was only used for external validation. It contained 228 patients with I-IIIA lung tumours treated with radiation or concurrent chemoradiation. Clinical characteristics of patients in the UMCG and Maastro datasets are summarized in Table 1 and in the supplemental file.

Deep learning on image features

A 3D convolutional neural network mainly consisting of 4 residual blocks was implemented for OS prediction. Normally, the number of patients in the stage I-IIIA NSCLC patient cohort with a death event at 2 years is relatively low (25%) [12]. Therefore, the deep learning model tends to predict no events due to the issue of imbalanced data. To tackle this problem, the focal loss function focusing more on the minorities than the binary cross entropy was used [13]. To provide a robust prediction, the average predicted value on cubes from CT scans of the same patient was calculated as the final prediction probability. Details in image processing and training of the deep learning model can be found in the supplementary file.

Machine learning on clinical factors

This study included the following clinical variables: age, sex (male vs female), T stage (T1 vs T2-T4), tumour stage (I vs II-IIIA) and the prescribed dose (as biologically effective-dose [BED]). Tumour stages were defined as stated in the seventh edition manual of American Joint Committee on Cancer [14]. The number of prescribed fractions, the fraction dose and the α/β ratio of 10 Gy were used for BED calculation, [15].

A clinical model was built using logistic regression for overall survival prediction at 2 years after treatment. Specifically, univariable analysis was first performed to preliminarily identify

prognostic features. The input feature with a p-value larger than 0.2 in the univariable analysis was excluded from further analysis [16]. After that, we used a bootstrap model selection method to perform 1000 times multivariable analysis with logistic regression using backward selection based on the likelihood ratio test [17]. This bootstrap method aimed to find a robust model with the most frequently selected features by the selection of remaining features repeated on 1000 bootstrap samples on the training set. At last, only the features selected more than 500 times in the bootstrap model selection method were considered as relevant variables to create the final logistic regression model for overall survival prediction.

Development of the hybrid model

We also designed a hybrid model that took both significant clinical features selected in the multivariable analysis and image features extracted by deep learning into account. The hybrid model illustrated in Fig. 1 was modified based on the deep learning model. In addition to the input of tumour cubes, previously selected clinical factors including continuous and categorical variables were set as the second input and concatenated with the image features after the global average pooling layer. The predicted overall survival probability was generated after the dense layer with the soft max function.

After predictions of the hybrid model, patients were stratified into low- and high-risk groups of mortality according to the optimal cut-off value of 0.3259 for the predicted mortality risk at 2 years that resulted in a maximum sum of sensitivity and specificity in the receiver operating characteristic curve on the UMCG training cohort.

Visualisation of the deep learning model

To gain a better insight into how deep learning inferred the survival likelihood for patients, the visualization method named Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to present an attention map showing where the deep learning focused on [18].

Statistical analysis

Statistical analysis was conducted using SPSS Statistics (version 22) and Python (version 3.7). P-values smaller than 0.05 were determined as statistically significant. Comparisons of continuous variables between training and test sets were analysed by t-test. Categorical variables in different sets were compared using chi-square comparisons or Fisher's exact test. When one of the entries was smaller than 10 in the frequency table, the Fisher's exact test ran. Otherwise, the chi-square test was applied. The Hosmer-Lemeshow test was used to evaluate the goodness-of-fit of the clinical model. Kaplan-Meier survival analysis and the log-rank test were used to compare the distributions between low- and high-risk groups. The performance of the prediction model was assessed using the area under the curve (AUC), sensitivity and specificity. The 95% confidence intervals (CI) for the AUCs were computed with DeLong's method [19], 95% CI for sensitivity and specificity were also reported according to the reference [20].

Results

The performance of the developed clinical model, deep learning model and hybrid model, on the independent UMCG and Maastro test sets, was shown in Table 2. The deep learning model achieved a median AUC of 0.62 [95% CI: 0.48-0.75] and 0.58 [95% CI: 0.51-0.64] on the complete UMCG and Maastro test sets (I-IIIA tumours), respectively. Regarding the clinical model, age and clinical stage were associated ($p = 0.053$ and $p = 0.053$, respectively) with OS in the univariable analysis. These two variables were also frequently selected in the

bootstrapping and showed statistical significance in the multivariable analysis ($p = 0.043$ for age and $p = 0.043$ for clinical stage). A logistic model was then developed and the Hosmer-Lemeshow test indicated the model fit the data well (p=0.899). The clinical model had similar AUCs as the deep learning model on the two complete test datasets. The hybrid model that combined prognostic clinical and image features presented a better median AUC of 0.76 [95% CI: 0.65-0.86] and 0.64 [95% CI: 0.58-0.70] on the two complete test sets than that of other models. Since the training UMCG set mainly consisted of tumours at stage I-II, sub-analyses were also performed on the Maastro test subset with only I-II tumours or IIIA tumours (supplementary file).

The Kaplan-Meier survival analysis was performed on the UMCG and Maastro test sets using the hybrid model that had the best AUCs. The Kaplan-Meier survival curves are shown in Fig. 2. The model showed discriminative separation between low and high mortality risk groups on the UMCG test set with ($p < 0.001$). Similarly, a p-value of 0.012 on the complete Maastro dataset indicated significant differences between two groups for overall survival at 2 years. The areas in the images that impacts survival prediction are visualized in Fig. 3. The most important regions were highlighted in red, while blue zones were relatively irrelevant. We found the deep learning model was mainly concentrated on partial tumour regions and these regions could contain prognostic patterns, such as lobulated shape. In addition, some regions surrounding the tumour were also enhanced, which may indicate tumour invasion. By contrast, areas with low intensity such as lung parenchyma showed fewer contributions to the prediction.

Discussion

In this study, we developed and validated the prognostic significance of a hybrid deep learning-based model to predict 2-year overall survival for stage I-IIIA NSCLC patients. By integrating clinical variables and image features on pre-treatment CT scans, the hybrid model had

reasonable prognostic performance with a median AUC of 0.76 [95% CI: 0.65-0.86] and 0.64 [95% CI: 0.58-0.70] on the complete UMCG and Maastro test sets, outperforming the clinical and deep learning models . We also demonstrated the capability of the hybrid model in stratifying patients into low and high mortality risk groups.

Numerous studies attempted to use clinical factors for prediction of overall survival since these features are commonly used in the clinic. In the current study, we found that age was significantly associated with OS, which is supported by several other studies [21, 22]. T-stage was not identified as a prognostic factor, although it is associated with tumour size which has been shown to be a prognostic factor [5]. A possible explanation could be that most patients with a tumour at T2-T3 in the UMCG cohort did not have an event within 2 years. However, clinical stage showed prognostic value, which is in agreement with other studies [23, 24]. This indicates that the combination of tumour size (T-stage) and lymph node status are prognostic for OS. Our finding that BED_{10} was not a significant predictor for overall survival was in agreement with the results of Kang et al.[5]. Other studies showed that a higher BED might be more effective to control NSCLC for the larger tumours only (T2 (29) or > 11 cc (30)) and brings survival benefits for individuals [25, 26]. The LQ model used for BED calculations is only applicable in a limited range of fraction doses. Therefore, the BEDs of the very large fraction doses (>=12 Gy) could be overestimated somewhat in our study. However, the models appeared to work surprisingly well in stereotactic treatments of lung cancer patients with fraction doses up to 20 Gy [27].

The deep learning model achieved similar AUC values as the clinical model on both test sets, indicating the deep learning-based image features on pre-treatment CT scans had value for

prognosis. The visualization method (i.e., the attention maps) provided insight into deep learning image patterns which were relevant to the prediction.

After combining the features in clinical and deep learning models, the hybrid model had median AUCs of 0.76 [95% CI: 0.65-0.86] and 0.64 [95% CI: 0.58-0.70] which were higher than that of clinical model or deep learning model on the complete UMCG and Maastro test sets. Note that, the hybrid model could still perform reasonably well, even though patient characteristics were different in these two test sets. This showed that clinical features provided complementary information to the deep learning model, resulting in an improved performance of the hybrid model in OS prediction. The performance of the hybrid model was comparable to that of other studies focusing on prognosis of the tumours at stage I-III. Xu et al. [8] built a deep learning model utilizing not only pre-treatment CT scans but also post-treatment CT scans and that model had an AUC of 0.74 for 2-year overall survival (stage III tumours) on the hold-out dataset. This performance was achieved when using 1-, 3-, and 6-month follow-up scans. For clinical practice this approach may be less suitable because of difficulties in collecting and preparing these follow-up scans. In addition, follow-up information will not be available before commencing treatment so these models cannot be used for decision-making in treatment strategies, e.g., individualized dose intensification or de-intensification. By contrast, in our study, only the pre-treatment scans were required for modelling and fewer human resources are needed. More importantly, the prediction is available before start of treatment, which makes it possible to adjust the treatment strategy. Moreover, Hosny et al. [10] performed a model for 2-year overall survival prediction of patients with tumours at stage I-IIIB and their median AUC was 0.70 [95% CI: 0.63–0.78] on the external test set. The larger training dataset available (464 patients) may be the reason why the model had better median AUC than ours: models could learn image features better from more samples with diverse morphological characteristics.

Furthermore, a sub-analysis on the Maastro test subset was performed and the hybrid model had a slightly better a median AUC of 0.66 [95% CI: 0.58-0.75] on tumours at stage I-II rather than at stage IIIA with 0.61 [95% CI: 0.51-0.70] (supplementary file). The reason could be that the model was mainly trained on the I-II tumours and image features were mainly learned from these tumours. Therefore, the model tended to have better prediction for overall survival on I-II tumours than I-IIIA tumours.

The present study has some limitations. First, we considered five clinical factors to build the clinical model. Other variables, such as Eastern Cooperative Oncology Group performance status score, smoking status, Charlson comorbidity index have shown statistical significance in OS prediction but were not available for this study [4, 5]. It could be interesting to include these variables to further improve the performance of the clinical and hybrid models. Second, the hybrid model has prognostic value in identifying patients with low and high mortality, but the predictive value in terms of discrimination in the locally advanced cases is relatively poor. It is expected that the model performance in the locally advanced cases can be improved by adding more advanced tumours in the training set. More research is necessary to investigate the options to intensify or de-intensify the treatment for the patient groups with high and low mortality risk. Third, there was heterogeneity between the UMCG training cohort and Maastro test set at the tumour level. The Maastro test set contained more patients with stage III disease (see table 1) compared to the UMCG training set. Stage III tumours have worse survival rates than the tumours at a lower stage. This could explain that the survival rates of the Maastro patients at low risk were lower than that of the UMCG patients at low risk in figure 2. Due to the difference in tumour size, the developed hybrid model might overestimate the survival rate of patients in the Maastro test set. Nevertheless, it is challenging to find an appropriate external test dataset

with the same tumour distributions and treatment methods for validation. Therefore, the evaluation on more external test datasets is required before applying it in clinical practice.

In conclusion, we presented a hybrid deep learning-based model that considered both clinical and image features for overall survival prediction at two years in stage I-IIIA NSCLC patients who received radiation therapy. Such a hybrid model can be utilized to identify patients with worse survival outcomes and guide clinical decision making in order to optimize their treatment.

References

[1] Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. 2019;69:363-85.
[2] Foster CC, Sher DJ, Rusthoven CG, Verma V, Spiotto MT, Weichselbaum RR, et al. Overall survival according to immunotherapy and radiation treatment for metastatic non-small-cell lung cancer: a National Cancer Database analysis. Radiation Oncology. 2019;14:1-13.
[3] Wrona A, Mornex F. Hypofractionation in Early Stage Non-Small Cell Lung Cancer. Seminars in Radiation Oncology2021. p. 97-104.
[4] Louie AV, Haasbeek CJ, Mokhles S, Rodrigues GB, Stephans KL, Lagerwaard FJ, et al. Predicting overall survival after stereotactic ablative radiation therapy in early-stage lung cancer: development and external validation of the Amsterdam prognostic model. Int J Radiat Oncol Biol Phys. 2015;93:82-90.
[5] Kang J, Ning MS, Feng H, Li H, Bahig H, Brooks ED, et al. Predicting 5-Year Progression and Survival Outcomes for Early Stage Non-small Cell Lung Cancer Treated with Stereotactic Ablative Radiation Therapy: Development and Validation of Robust Prognostic Nomograms. Int J Radiat Oncol Biol Phys. 2020;106:90-9.
[6] Zheng S, Cornelissen LJ, Cui X, Jing X, Veldhuis RN, Oudkerk M, et al. Deep convolutional neural networks for multi‐planar lung nodule detection: improvement in small nodule identification. Med Phys. 2020;48:733-44.
[7] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25:954-61.
[8] Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin Cancer Res. 2019;25:3266-75.
[9] Mukherjee P, Zhou M, Lee E, Schicht A, Balagurunathan Y, Napel S, et al. A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets. Nat Mach Intell. 2020;2:274-82.
[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] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:1-9.
[12] Chiang A, Thibault I, Warner A, Rodrigues G, Palma D, Soliman H, et al. A comparison between accelerated hypofractionation and stereotactic ablative radiotherapy (SABR) for early-stage non-small cell lung cancer (NSCLC): Results of a propensity score-matched analysis. Radiother Oncol. 2016;118:478-84.
[13] Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision2017. p. 2980-8.
[14] Edge SB, Compton CCJAoso. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17:1471-4.
[15] Santiago A, Barczyk S, Jelen U, Engenhart-Cabillic R, Wittig AJRO. Challenges in radiobiological modeling: can we decide between LQ and LQ-L models based on reviewed clinical NSCLC treatment outcome data? Radiat Oncol. 2016;11:1-13.
[16] Irie M, Nakanishi R, Yasuda M, Fujino Y, Hamada K, Hyodo MJERJ. Risk factors for short-term outcomes after thoracoscopic lobectomy for lung cancer. Eur Respir J. 2016;48:495-503.
[17] Austin PC, Tu JVJTAS. Bootstrap methods for developing predictive models. The American Statistician. 2004;58:131-7.
[18] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision2017. p. 618-26.
[19] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12:1-8.
[20] Zhu W, Zeng N, Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland. 2010;19:67.
[21] Zheng X, Sun Y, Ye K, Fan C, Wang X, Yang Y, et al. Stereotactic ablative radiotherapy as single treatment for early stage non‐small cell lung cancer: A single institution analysis. Thorac Cancer. 2021.
[22] Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. Int J Radiat Oncol Biol Phys. 2017;99:344-52.
[23] Mokhles S, Nuyttens JJ, Maat AP, Birim Ö, Aerts JG, Bogers AJ, et al. Survival and treatment of non-small cell lung cancer stage I–II treated surgically or with stereotactic body radiotherapy: patient and tumor-specific factors affect the prognosis. Ann Surg Oncol. 2015;22:316-23.
[24] Liao Y, Wang X, Zhong P, Yin G, Fan X, Huang CJOl. A nomogram for the prediction of overall survival in patients with stage II and III non‐small cell lung cancer using a population‐based study. Oncol Lett. 2019;18:5905-16.
[25] Koshy M, Malik R, Weichselbaum RR, Sher DJJIJoROBP. Increasing radiation therapy dose is associated with improved survival in patients undergoing stereotactic body radiation therapy for stage I non–small-cell lung cancer. Int J Radiat Oncol Biol Phys. 2015;91:344-50.
[26] Suzuki O, Mitsuyoshi T, Miyazaki M, Teshima T, Nishiyama K, Ubbels JF, et al. Dose–volume–response analysis in stereotactic radiotherapy for early lung cancer. Radiother Oncol. 2014;112:262-6.
[27] McMahon SJJPiM, Biology. The linear quadratic model: usage, interpretation and challenges. Physics in Medicine & Biology. 2018;64.

Tables

Table 1. Clinical characteristics of patients in the UMCG and Maastro datasets

Characteristics	UMCG training set (n=189)	UMCG test set1 (n=81)	P-value (training set vs test set1)	Maastro test set2 (n=228)	P-value (training set vs test set2)
Age, Median (range)	76 (55-91)	73 (49-88)	0.060	71 (33-91)	<0.001
Sex (%)			0.594		0.005
Male	106 (56.1)	42 (51.9)		158 (69.3)	
Female	83 (43.9)	39 (48.1)		70 (30.7)	
T stage (%)			0.438		<0.001
T1	149 (78.8)	64 (79.0)		71 (31.1)	
T2	34 (18.0)	12 (14.8)		112 (49.1)	
T3	5 (2.6)	3 (3.7)		45 (19.7)	
T4	1 (0.5)	2 (2.5)		0 (0.0)	
Clinical stage (%)			0.627		<0.001
I	177 (93.7)	74 (91.4)		84 (36.8)	
II	10 (5.3)	5 (6.2)		36 (15.8)	
IIIA	2 (1.1)	2 (2.5)		108 (47.4)	
BED_{10} (Gy), Median (range)	132 (68.4-168.0)	132 (85.5-187.5)	0.169	-	-
2-year OS (class)			0.927		<0.001
Survival	141	60		89	
Death	48	21		139	

Table 2. Performance of the diverse models evaluated on independent UMCG and Maastro test sets for the predictive risk of overall survival at two years after treatment. Results are presented using a 95% confidence interval.

	AUC	Sensitivity	Specificity
Deep learning model			
UMCG training set (I-IIIA tumours)	0.90 [0.86-0.94]	0.81 [0.68-0.90]	0.87 [0.80-0.91]
UMCG test set (I-IIIA tumours)	0.62 [0.48-0.75]	0.33 [0.17-0.55]	0.92 [0.82-0.96]
UMCG complete set (I-IIIA tumours)	0.82 [0.77-0.87]	0.71 [0.59-0.80]	0.82 [0.76-0.87]
Maastro test set (I-IIIA tumours)	0.58 [0.51-0.64]	0.65 [0.57-0.72]	0.55 [0.45-0.65]
Clinical model			
UMCG training set (I-IIIA tumours)	0.61 [0.52-0.71]	0.48 [0.34-0.62]	0.66 [0.58-0.73]
UMCG test set (I-IIIA tumours)	0.62 [0.50-0.76]	0.71 [0.50-0.86]	0.45 [0.33-0.58]
UMCG complete set (I-IIIA tumours)	0.62 [0.54-0.69]	0.72 [0.61-0.82]	0.38 [0.32-0.45]
Maastro test set (I-IIIA tumours)	0.57 [0.50-0.65]	0.76 [0.68-0.82]	0.36 [0.27-0.46]
Hybrid model			
UMCG training set (I-IIIA tumours)	0.83 [0.78-0.88]	0.73 [0.59-0.83]	0.78 [0.70-0.84]
UMCG test set (I-IIIA tumours)	0.76 [0.65-0.86]	0.62 [0.41-0.79]	0.80 [0.68-0.88]
UMCG complete set (I-IIIA tumours)	0.80 [0.74-0.84]	0.65 [0.53-0.75]	0.82 [0.76-0.86]
Maastro test set (I-IIIA tumours)	0.64 [0.58-0.70]	0.76 [0.69-0.83]	0.52 [0.41-0.62]

Figures

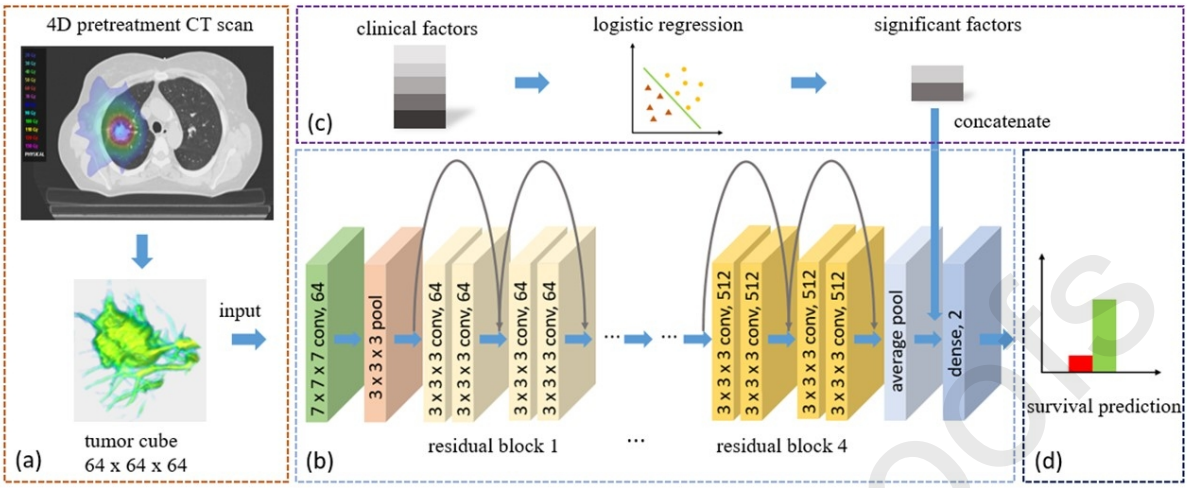

Fig. 1. Illustration of the hybrid model. (a) Cubic patches including tumours were extracted from the 4D pre-treatment CT scans. (b) The architecture of the 3D convolutional neural: each of the four residual blocks contained four convolutional layers (only block 1 and 4 are shown). The number of filters doubled after every block. Grey arrows indicate shortcut connections. (c) The prognostic clinical factors were selected using logistic regression. These variables were concatenated with image features extracted by deep learning in the hybrid model. (d) The survival likelihood was generated after the dense layer.

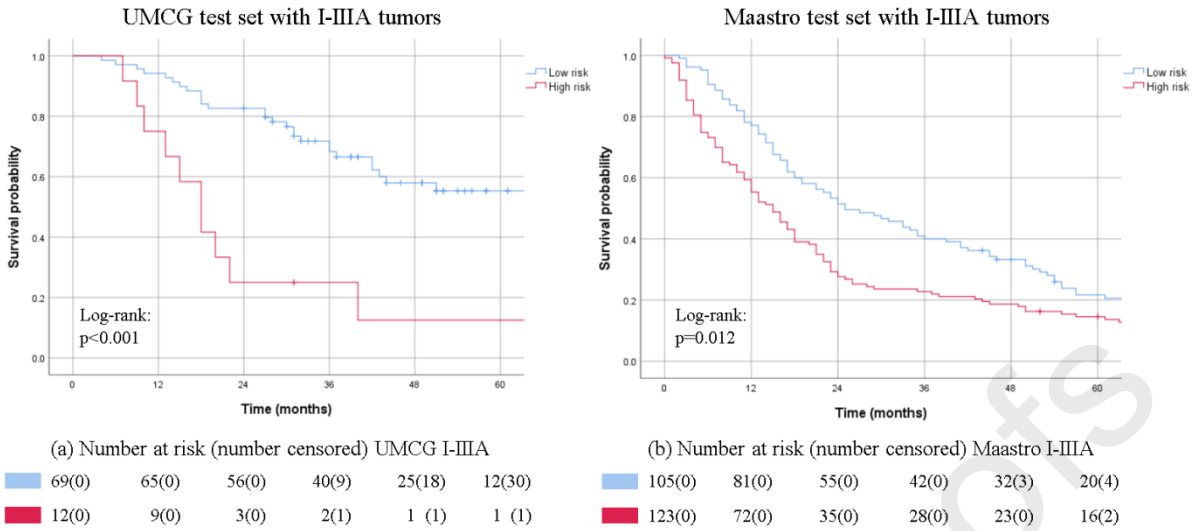

Fig. 2. Kaplan Meier curves for overall survival in independent UMCG and Maastro test cohorts for the low- and high-risk groups. The optimal cut off value, used to stratify patients into low and high mortality risk groups, was calculated in the UMCG training cohort, based on the predicted morality risk at 2 years that maximized the sum of sensitivity and specificity in the ROC curve.

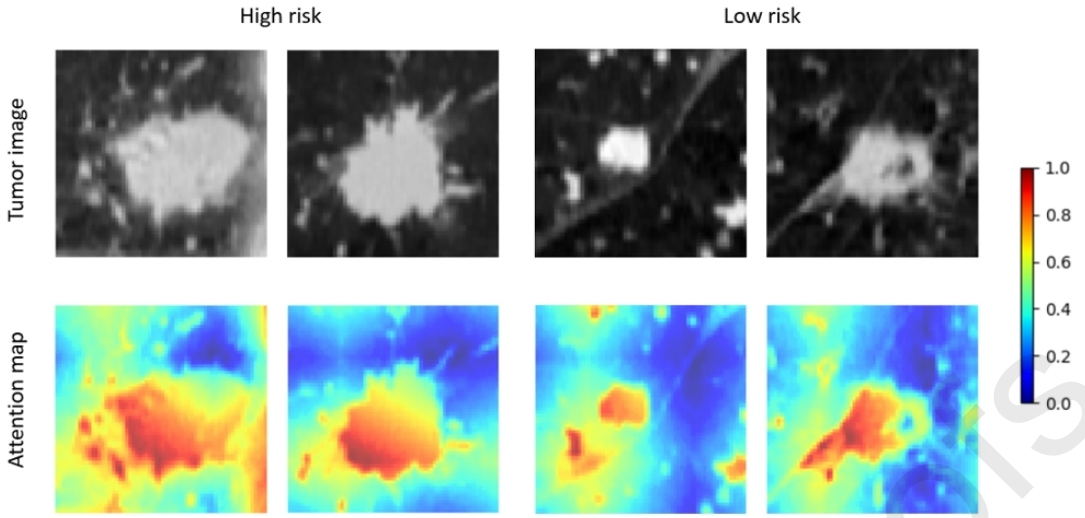

Fig. 3. Visualization of the attention map for survival prediction: red indicates the regions which contain the most prognostic image features.

Highlights

1. By considering clinical variables and image features from pre-treatment CT-scans, a hybrid deep learning-based model was implemented for overall survival prediction at two years in stage I-IIIA NSCLC patients.
2. The proposed model achieved reasonable performance on different patient groups.
3. The developed hybrid model has the potential to identify patients with high mortality risk and guide clinical decision making.