A deep learning-based system for survival benefit prediction of tyrosine kinase inhibitors and immune checkpoint inhibitors in stage IV non-small cell lung cancer patients: A multicenter, prognostic study

Kexue Deng,[a,1,**] Lu Wang,[b,c,1] Yuchan Liu,[a] Xin Li,[d] Qiuyang Hou,[a] Mulan Cao,[d] Nathan Norton Ng,[e] Huan Wang,[f] Huanhuan Chen,[g] Kristen W. Yeom,[e] Mingfang Zhao,[d] Ning Wu,[h] Peng Gao,[i,**] Jingyun Shi,[j,**] Zaiyi Liu,[k,l,**] Weimin Li,[m,**] Jie Tian,[n,o,p,q,**] and Jiangdian Song [c,*]

[a]Department of radiology, The First Affiliated Hospital of University of Science and Technology of China (USTC), Division of Life Sciences and Medicine, USTC, Hefei, Anhui, China
[b]Library of Shengjing Hospital of China Medical University, Shenyang, China
[c]School of Health Management, China Medical University, Shenyang, Liaoning, China
[d]Department of Medical Oncology, The First Hospital of China Medical University, Shenyang, Liaoning, China
[e]Department of Radiology, School of Medicine Stanford University, Stanford CA 94305, United States
[f]Radiation oncology department of thoracic cancer, Liaoning Cancer Hospital and Institute, Liaoning, China
[g]Department of Oncology, Shengjing Hospital of China Medical University, Shenyang, China
[h]PET-CT center, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[i]Department of Surgical Oncology and General Surgery, The First Affiliated Hospital of China Medical University, Shenyang, Liaoning, China
[j]Department of Radiology, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China
[k]Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China
[l]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China
[m]Department of Respiratory and Critical Care Medicine, West China Hospital, Chengdu, Sichuan, China
[n]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing, China
[o]Key Laboratory of Big Data-Based Precision Medicine, Beihang University, Ministry of Industry and Information Technology, Beijing, China
[p]Engineering Research Center of Molecular and Neuro Imaging of Ministry of Education, School of Life Science and Technology, Xidian University, Xi'an, China
[q]CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging, Beijing, China

Summary

Background For clinical decision making, it is crucial to identify patients with stage IV non-small cell lung cancer (NSCLC) who may benefit from tyrosine kinase inhibitors (TKIs) and immune checkpoint inhibitors (ICIs). In this study, a deep learning-based system was designed and validated using pre-therapy computed tomography (CT) images to predict the survival benefits of EGFR-TKIs and ICIs in stage IV NSCLC patients.

Methods This retrospective study collected data from 570 patients with stage IV EGFR-mutant NSCLC treated with EGFR-TKIs at five institutions between 2010 and 2021 (data of 314 patients were from a previously registered study), and 129 patients with stage IV NSCLC treated with ICIs at three institutions between 2017 and 2021 to build the ICI test dataset. Five-fold cross-validation was applied to divide the EGFR-TKI-treated patients from four institutions into training and internal validation datasets randomly in a ratio of 80%:20%, and the data from another institution was used as an external test dataset. An EfficientNetV2-based survival benefit prognosis (ESBP) system was developed with pre-therapy CT images as the input and the probability score as the output to identify which patients would receive additional survival benefit longer than the median PFS. Its prognostic performance was validated on

eClinicalMedicine 2022;51: 101541
Published online xxx
https://doi.org/10.1016/j.eclinm.2022.101541

*Corresponding author at: School of Health Management, China Medical University, Shenyang, Liaoning 110122, China.
E-mail address: song.jd0910@gmail.com (J. Song).
[1] These authors contributed equally to this work.
** Co-corresponding authors.

the ICI test dataset. For diagnosing which patient would receive additional survival benefit, the accuracy of ESBP was compared with the estimations of three radiologists and three oncologists with varying degrees of expertise (two, five, and ten years). Improvements in the clinicians' diagnostic accuracy with ESBP assistance were then quantified.

Findings ESBP achieved positive predictive values of 80·40%, 75·40%, and 77·43% for additional EGFR-TKI survival benefit prediction using the probability score of 0·2 as the threshold on the training, internal validation, and external test datasets, respectively. The higher ESBP score (>0·2) indicated a better prognosis for progression-free survival (hazard ratio: 0·36, 95% CI: 0·19−0·68, $p<0·0001$) in patients on the external test dataset. Patients with scores >0·2 in the ICI test dataset also showed better survival benefit (hazard ratio: 0·33, 95% CI: 0·18−0·55, $p<0·0001$). This suggests the potential of ESBP to identify the two subgroups of benefiting patients by decoding the commonalities from pre-therapy CT images (stage IV EGFR-mutant NSCLC patients receiving additional survival benefit from EGFR-TKIs and stage IV NSCLC patients receiving additional survival benefit from ICIs). ESBP assistance improved the diagnostic accuracy of the clinicians with two years of experience from 47·91% to 66·32%, and the clinicians with five years of experience from 53·12% to 61·41%.

Interpretation This study developed and externally validated a preoperative CT image-based deep learning model to predict the survival benefits of EGFR-TKI and ICI therapies in stage IV NSCLC patients, which will facilitate optimized and individualized treatment strategies.

Funding This study received funding from the National Natural Science Foundation of China (82001904, 81930053, and 62027901), and Key-Area Research and Development Program of Guangdong Province (2021B0101420005).

Keywords: Non-small cell lung cancer; Tyrosine kinase inhibitor; Immune checkpoint inhibitor; Artificial intelligence; Survival benefits

Research in context

Evidence before this study

We searched PubMed and Google Scholar on November 21, 2021, using the keywords: "artificial intelligence" OR "deep learning" OR "machine learning" AND "EGFR-TKI" OR "ICI" AND "non-small cell lung cancer", yielded 450 results including 25 original studies that employed data analysis to predict the survival of EGFR-TKIs or ICIs. However, only one study explores the potential of deep learning models for the prognosis of both EGFR-TKI and ICI treatments using pre-therapy radiological images. None of these studies compared diagnosis between the models and clinicians, nor evaluated the diagnostic accuracy improvement to clinicians when assisted by a deep learning model for the prognosis of stage IV NSCLC.

Added value of this study

To the authors' knowledge, this is the first study to evaluate improvement in diagnoses of radiologists and oncologists with different levels of expertise of such survival benefits when assisted by a deep learning model. The proposed ESBP is confirmed as a prognostic tool for EGFR-TKI and ICI therapy, which suggests the potential of ESBP to identify the two subgroups of benefiting patients by decoding the commonalities from pre-therapy CT images. With the assistance of ESBP, the diagnostic accuracy of the clinicians with two and five years of experience regarding the survival benefit from EGFR-TKI treatment increased by 18·14% and 8·29%, respectively, reaching a level comparable to the diagnostic accuracy of expert clinicians.

Implications of all the available evidence

The proposed model, which requires no additional training of personnel, is clinically applicable to predict the survival benefit of stage IV NSCLC patients and improves the performance of non-expert radiologists and oncologists to the level of experts. The ESBP model may have value as an automated screening tool to triage stage IV NSCLC patients whose EGFR-TKIs and ICIs survival benefit status is uncertain, thus potentially improving the probability of treatment benefit and increasing the efficiency of treatment-related labor and costs.

Introduction

Lung cancer is the second-most commonly diagnosed malignancy and the leading cause of cancer-related deaths worldwide.[1] Non-small cell lung cancer (NSCLC) accounts for 80−85% of all lung cancer cases.[2] It is

recommended that stage IV NSCLC patients who are unable to undergo surgical resection be tested to determine their epidermal growth factor receptor (EGFR) status and PD-L1 expression. A positive EGFR mutation identifies a distinct NSCLC patient subgroup with a better prognosis, in whom EGFR tyrosine kinase inhibitors (TKIs) are critical for prolonging life,[3–5] and immune checkpoint inhibitors (ICIs) significantly improve the survival benefit of patients with PD-L1 expression.[6]

Previous studies have reported a treatment response observed in 70% of EGFR-mutant NSCLC patients following the clinical administration of EGFR-TKI drugs, whereas the objective response rate of ICIs in NSCLC patients was only ≤45%.[4,7,8] In clinical practices, EGFR mutations occur in 40–60% of patients with NSCLC, which is much higher than in patients with PD-L1 expression and other mutation subtypes.[9] Although Osimertinib has been demonstrated to be effective against acquired TKI resistance caused by secondary T790M mutations, the risk of treatment failure remains uncertain, because definite biomarkers, that identify patients who would receive an additional survival benefit from EGFR-TKI, are rare.[10] To date, the median progression-free survival (PFS) of EGFR-TKI responders is approximately 9·5 months.[11] With the increasing clinical usage of ICI therapies, it is critical to develop computer-assisted prognostic tools to aid patient selection for EGFR-TKI and ICI administration; this would help target patients who have a high probability of benefiting from these drugs while reducing the labor and expense associated with treating patients who may not benefit from them.

Artificial intelligence (AI), particularly deep learning, has already shown potential for assisting in NSCLC treatment.[12,13] For example, AI can automatically extrapolate the subtle heterogeneity hidden in computed tomography (CT) images and identify latent semantics that are often undetectable by the human eye.[14,15] Two recent studies indicate that the deep learning systems for predicting EGFR mutation status inform the efficacy of evaluating EGFR-TKI and ICI therapy.[16,17] Although encouraging preliminary results have been published regarding the use of AI in the prognosis of stage IV NSCLC patients, studies are rare that compare the prognostic accuracy of deep learning models for the prediction of EGFR-TKI survival benefit with that of radiologists and oncologists with different levels of expertise. Even rarer are studies that evaluate the improvement of the clinicians' diagnoses of EGFR-TKI survival benefit with the assistance of deep learning models.

The EfficientNet architectures, especially the state-of-the-art EfficientNetV2, have demonstrated to assist the diagnoses of COVID-19 and other diseases via transfer learning.[18–21] This study was aimed at training and independently validating an EfficientNetV2-based survival benefit prognosis (ESBP) system to evaluate the survival benefits for both EGFR-TKI and ICI in stage IV NSCLC on pre-therapy CT images. The performance of ESBP was compared with that of radiologists and oncologists at three expertise levels (trainee, competent, and expert) on the same test set to assess whether ESBP added value to the current stage IV NSCLC clinical treatment paradigms.

Methods

Study design and participants

A retrospective multicenter study was conducted in China to collect stage IV EGFR-mutant NSCLC patients who received EGFR-TKI therapy at five independent institutions between January 1, 2010, and June 30, 2021 (Shanghai Pulmonary Hospital, Guangdong Provincial People's Hospital, West China Hospital, and two centers of The First Affiliated Hospital of University of Science and Technology of China), and stage IV NSCLC patients treated with ICI therapy at three independent institutions between January 1, 2017, and June 30, 2021 (Shengjing Hospital of China Medical University, Liaoning Cancer Hospital and Institute, and The First Affiliated Hospital of University of Science and Technology of China). In total, 570 EGFR-TKI-treated stage IV EGFR-mutant NSCLC patients and 129 ICI-treated stage IV NSCLC patients were enrolled in this study. The study's workflow is presented in Figure 1. The EGFR-TKI datasets were of the patients with positive EGFR mutation who underwent the recommended EGFR-TKI therapy (including the first to third generation TKI drugs). Among them, three hundred and ninety-eight patients received the EGFR-TKI drugs as their first-line treatment (21 censored), and 158 received it as their second-line treatment (49 censored). The third-generation TKI drug Osimertinib was administered to 33 patients, whereas 385 and 152 patients, received first-generation and second-generation EGFR-TKI drugs, respectively. The ICI test dataset consisted of patients who were recommended for ICI therapy and received the appropriate ICI treatment regimen. Among them, 19 patients with PD-L1 expression of higher than 50% were recruited, and 48 of whom received immunotherapy combined with chemotherapy.

We acquired the pre-therapy CT scans from the local PACS (Picture Archiving and Communication System) and the demographic data from each participating hospital. This study was approved by the institution's ethics committee, and informed consent was waived due to the retrospective nature of the study. Diagnostic CT scans taken within one month before drug administration were required for all eligible patients. Demographic information, including sex, age, smoking status, performance status score, histopathological subtype, EGFR mutation subtype (EGFR-TKI patients), tumor proportion score (ICI-treated patients), and the administered

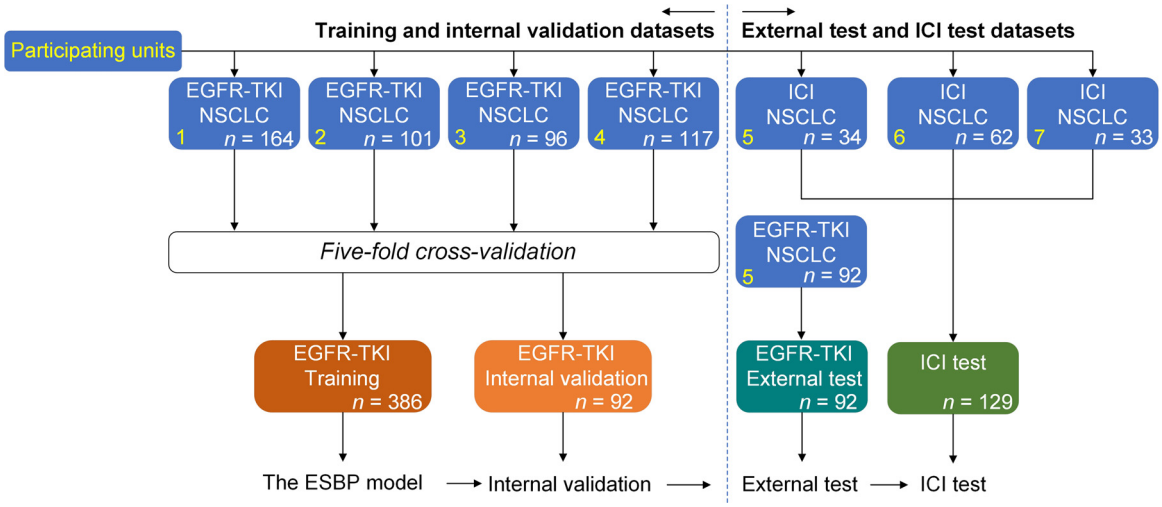

Figure 1. Workflow of this study.

therapeutic regimen, was required. Patients who had a history of systemic anticancer therapy for advanced disease, resection, or missing follow-up records were excluded.

This study used progression-free survival (PFS), measured from the date of drug administration to the time of disease progression by RECIST 1.1 criteria (at least a 20% increase in the sum of diameters of target lesions, or the appearance of one or more new lesions)[22] or the time of death, whichever occurred earlier. PFS was used to measure additional survival benefit, and the prognostic performance of ESBP for evaluating the additional survival benefit were then quantified. Patients who were alive without progression records were censored at the date of the last follow-up.[23] The analysis focused on PFS instead of overall survival because, in patients with stage IV NSCLC, the evaluation of overall survival could be affected by confounding factors such as other treatments after the EGFR-TKI/ICI treatment mentioned in this retrospective study.[24]

Based on the follow-up, all EGFR-TKI patients enrolled in this study with distinct disease progression events according to RESIST 1.1 criteria were screened to determine their PFS, and a median PFS of 9·5 months was determined. Blumenthal et al. confirmed a median PFS of approximately 9·5 months in the responders of EGFR-TKI therapies based on the review of 14 trials of targeted therapies.[11] In addition, median PFS has proven to be an effective cut-off value to classify the sensitivity or insensitivity to targeted therapy in NSCLC patients.[25] Consequently, the median PFS of 9·5 months was used as the cut-off value to define a patient who would obtain additional EGFR-TKI survival benefit. Patients classified into the low-risk group (good responders), whose PFS was higher than the median, were considered to receive additional survival benefit, and the high-risk (poor responders) group, whose PFS was lower than the median, were excluded from receiving additional survival benefit through EGR-TKI therapy.

Input data preparation

Local radiologists reviewed and delineated the regions of interest of the included NSCLC on CT images. The entire primary tumor, and at least 10 mm of its peripheral region, were included and then resized to 224 × 224 pixels using spline interpolation. For central tumors, the peripheral region was obtained with the tumor as the center. Meanwhile, for adhesive tumors, the peripheral region was obtained by manually determining the direction of expansion of the tumor into the lung parenchyma. Data augmentation including translation, flipping, rotation, and zooming was used and each image was normalized before it was input into the model for training (as presented in Supplementary Section A1). To obtain a robust model for processing images from diverse sources and of varied quality, the parameters and protocols of the CT scans were not limited.

Model development and validation

As shown in Figure 1, five-fold cross-validation was used for the patients treated with EGFR-TKI therapy from four participating institutions to divide a training and an internal validation dataset randomly in an 80:20 ratio. To develop the ESBP classifier, a state-of-the-art EfficientNetV2 architecture,[20] pre-trained on ImageNet, was used and fine-tuned via transfer learning on the training dataset, which used two-dimensional images as input and output image-level probabilities for survival benefit diagnosis. Image-level probabilities were aggregated into a scan-level probability by considering the mean of the CT slices including the primary tumor(s), which, when a patient had more than one pre-therapy

scan analyzed, was further averaged to obtain the patient-level score for survival benefit diagnosis. A threshold of the patient-level scores was subsequently selected using X-tile[26] software to dichotomize the patient data into subgroups that received additional survival benefit or not. Based on the threshold, the diagnostic accuracies on the training and internal validation datasets were obtained. The final ESBP model and the threshold were determined when the optimal diagnostic accuracy was achieved on the internal validation dataset. The final ESBP model was then validated on the external test dataset consisting of EGFR-TKI-treated patients from the other one participating institution. To further verify the ESBP model's performance, it was applied to the ICI test dataset to test its prognostic performance in ICI-treated stage IV NSCLC patients. The ESBP model is implemented using PyTorch (Version: 1.7.1), and the training details and hyper-parameters for this model are presented in Supplementary Section A1.

In addition, to compare the survival benefits of the poor and good responders by the ESBP with that of the NSCLC patients undergoing chemotherapy, advanced-stage NSCLC patients undergoing first-line chemotherapy were enrolled from three participating institutions. The enrollment details of the chemotherapy patients are presented in Supplementary Section A2.

Accuracy comparison and reproducibility analysis

A reader study was conducted to compare the performance of ESBP on EGFR-TKI survival benefit prediction with that of clinicians and to evaluate its impact on their performance. Three radiologists and three oncologists, with varying degrees of expertise, reviewed the same external test dataset. The two experts were professors with more than ten years of experience in radiology and oncology; the competent group consisted of two attending doctors with five years of experience in each field; and two trainee residents with two years of experience in each field.

All the clinicians were blinded for the follow-up, and the images were de-identified for their assessments. For each patient, the pre-therapy CT scan and clinical information, including sex, age, smoking history, histopathological subtype, and TNM stage, were provided to ensure that the clinicians' diagnostic procedures were consistent with the actual clinical workflow. All the clinicians recorded their diagnoses on whether the patient would receive additional survival benefit from EGFR-TKI therapy (i.e., assessing whether the PFS could be higher than the median of 9·5 months). Improvements to the clinicians' diagnosis were determined by comparing the first round of diagnosis without ESBP assistance, and the second round of ESBP assistance with the diagnostic result after a 4-week washout. The predicted survival benefit score (in continuous values) for each patient from ESBP, the final dichotomized prediction (by the threshold of X-tile) based on the scores, and the overall accuracy were presented to the clinicians for their second-round diagnostic decision. At the study's conclusion, the clinicians were asked whether, if available, they would use an artificial intelligence-based survival prediction tool to triage patients for the clinical diagnosis of EGFR-TKI survival benefit.

The manual segmentation of tumors in CT images requires professional expertise in delineating the tumor boundaries, a process that is highly susceptible to subjective experience. Therefore, an inter-observer reproducibility experiment was conducted to demonstrate that the proposed ESBP model is insensitive to variations in tumor boundaries delineated by different observers. Accurate manual delineation of the primary lung tumors was not required in this experiment and two other local radiologists were only asked to draw the approximate regions of the lung tumors. The example of tumor segmentation is presented in Supplementary Figure S3. To compare the scores for the accurately segmented tumor images, all image processing procedures were identical to those described earlier, and the images were fed into the proposed ESBP model to obtain the patient-level scores.

In addition, a previous study indicated that to predict EGFR mutation status, the images of lungs containing lesions produced better performance than the images of lung tumors.[27] Therefore, an ad-hoc comparative experiment was conducted using the lung images containing the primary tumor to train and test an ESBP model. The lung region was automatically segmented using an adaptive region-growing algorithm.[28] The EfficientNetV2 architecture and hardware conditions for train the two models were identical. Further details are presented in Supplementary Section A3.

Statistical analysis

For prognostic stratification, the ESBP scores were used to divide the patients into two subgroups for which the median PFS and hazard ratio (HR) were calculated, and the Cox proportional hazard assumption was evaluated. The prognostic performance of ESBP was validated using a Kaplan–Meier survival analysis, a log-rank test, an HR with 95% confidence level (CI), and a Harrell's concordance index (C-index) with 95% CI. Sample size evaluation was performed using PASS (Version: 21.0.3, NCSS, LLC, UTAH).[29]

For diagnostic classification, a time-dependent receiver operating characteristic (ROC) analysis was performed with 95% Wald CIs using the continuous ESBP score. In addition, a Kaplan–Meier survival analysis, a log-rank test, and an HR with 95% CI were used to evaluate the benefit difference between the two subgroups classified by each clinician.

For the reader study, the benefit prediction agreement between the two clinicians at each level when unassisted by ESBP was calculated using Fleiss' κ.[30] A McNemar's test[31] was used to evaluate the effect of ESBP on the

clinicians' predictions by comparing their diagnostic accuracy with and without ESBP assistance. In addition, Fleiss' κ was used to evaluate the variation in the survival benefit prediction of ESBP using the segmentation images of different radiologists on the same dataset.

Role of the funding source

The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The corresponding authors confirm that they had full access to the data and had the final responsibility for deciding to submit the manuscript for publication.

Results

A summary of the demographic variables and clinical characteristics of the EGFR-TKI training, internal validation, and external test datasets, and the ICI test dataset is provided in Table 1. There were no statistically significant differences in age, sex, smoking history, histopathological subtype, EGFR mutation subtype, PFS, or the administered therapeutic regimen between the EGFR-TKI datasets ($P > 0.05$).

Based on the result of five-fold cross-validation, the ESBP model was obtained when the best accuracy was achieved on the internal validation dataset. The subgroups were defined using an ESBP threshold of 0·2 calculated by X-tile, above which patients were predicted to be low-risk. Such results indicate improved survival benefit (good responders), whereas those below were classified as at high risk of not receiving the expected survival benefit (poor responders). The results indicated that the ESBP classifier was a strong predictor of PFS in the primary analysis of the 386 patients in the EGFR-TKI training dataset [for high-risk (187 patients a with median PFS of 6·4 months) vs. low-risk (199 patients with a median PFS of 18·2 months), HR: 5·07, 95% CI: 3·80−6·78, $P < 0.0001$], the EGFR-TKI internal validation dataset [for high-risk (31 patients with a median PFS of 5·0 months) vs. low-risk (61 patients with a median PFS of 17·5 months), HR: 3·50, 95% CI: 1·76 −6·95, $P < 0.0001$], and the EGFR-TKI external test dataset [for high-risk (28 patients with a median PFS of 7·2 months) vs. low-risk (64 patients with a median PFS of 17·6 months), HR: 2·77, 95% CI: 1·50−5·30, $P < 0.0001$]. When using the ESBP threshold of 0·2 on the ICI test dataset, the higher scoring patients (91 cases with a median PFS of 7·0 months) showed significantly better survival benefit than the lower scoring patients (38 cases with a median PFS of 3·0 months, HR: 0·33, 95% CI: 0·18−0·55, $P < 0.0001$), as shown in Figure 2.

	EGFR-TKI				ICI	
	Training and internal validation ($N = 478$)		External test ($N = 92$)		ICI test ($N = 129$)	
	PFS < 9·5	*PFS* > 9·5	*PFS* < 9·5	*PFS* > 9·5	*PFS* < 9·5	*PFS* > 9·5
Number	249	229	36	56	41	88
Age, years (SD)	58 (2·4)	59 (5·9)	61 (9·5)	60 (7·5)	59 (8·2)	63 (5·5)
Sex						
Male	106	86	12	28	30	62
Female	143	143	24	28	11	26
Smoke (yes)	46	46	13	9	30	55
Pathology						
ADE	230	221	33	53	40	80
Others	19	8	3	3	1	8
EGFR mutation						
19-del	92	90	14	16	*NA*	*NA*
21L858R	77	70	11	11	*NA*	*NA*
Other	80	69	11	29	*NA*	*NA*
PD-L1 status						
≥50%	*NA*	*NA*	*NA*	*NA*	4	15
1%−49%	*NA*	*NA*	*NA*	*NA*	5	17
Other	*NA*	*NA*	*NA*	*NA*	32	56
Median PFS (SD)	6·0 (2·7)	15·9 (10·5)	5·5 (2·6)	16·6 (7·6)	4·0 (2·6)	13·6 (7·9)

***Table 1*: Baseline characteristics and PFS of the patients treated with EGFR-TKI enrolled from four hospitals to construct the EGFR-TKI training and internal validation datasets (five-fold cross-validation), and the patients to construct the EGFR-TKI external test dataset, and the ICI test dataset.**

Note: The data from the training and internal validation datasets are combined because five-fold cross-validation was performed to train and internal validate the ESBP model, and the data in the training and internal validation datasets are different in each fold.

ADE = Adenocarcinoma, SD: standard deviation, *NA*: not applicable. PFS is measured in months.

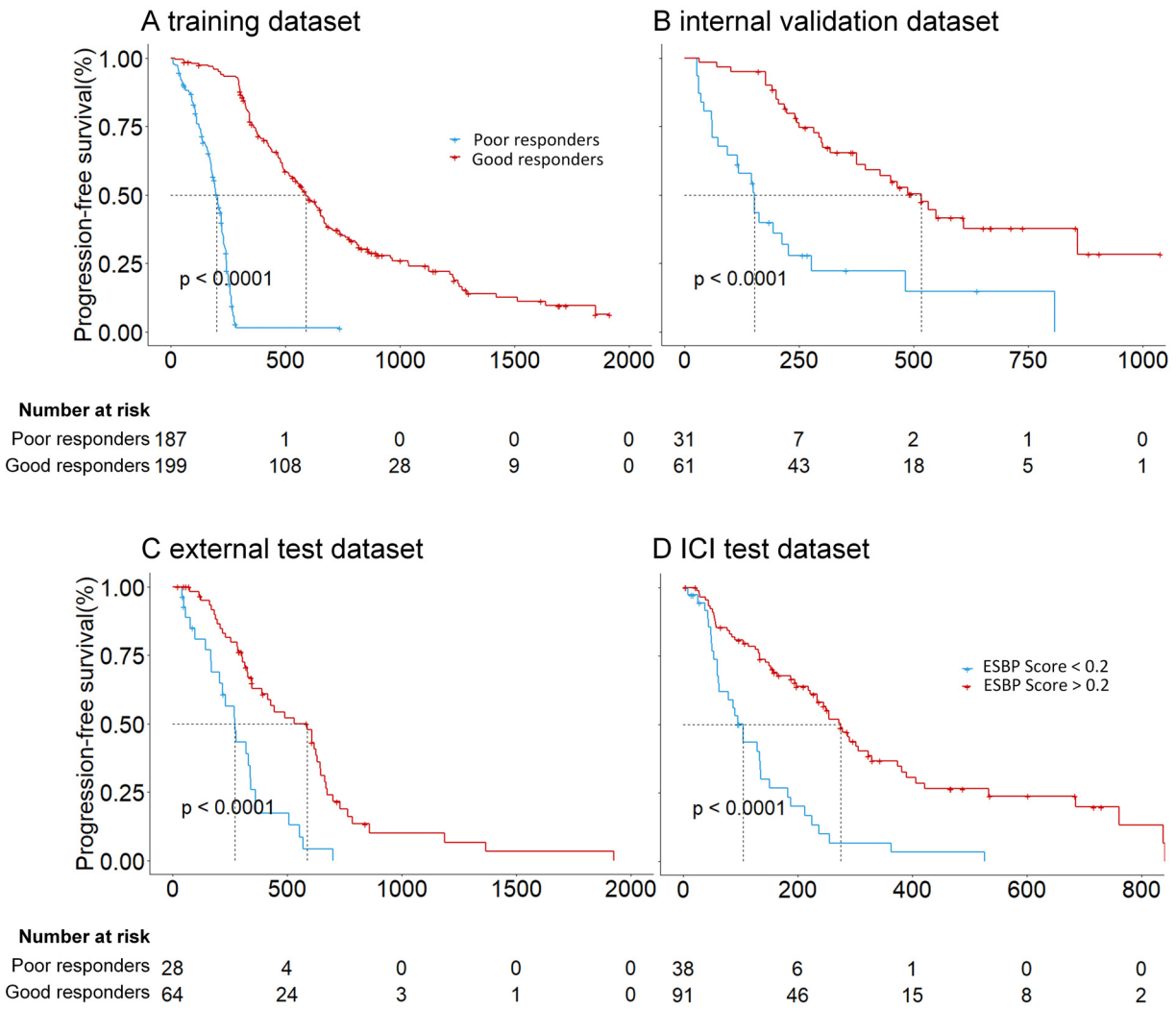

Figure 2. Kaplan–Meier analysis of progression-free survival by the EfficientNetV2-based survival benefit prediction system (ESBP) classifier, evaluated on training (A), internal validation (B), and external test (C) datasets, and further validated on the ICI test dataset (D).

Diagnosis accuracy	First diagnosis	Second diagnosis	*P* value (*McNemar's test*)
ESBP	76·08%	*NA*	*NA*
Radiologists			
Trainee	47·93%	68·50%	<0·0001
Competent	51·03%	60·86%	<0·0001
Expert	65·09%	70·70%	0·180
Oncologists			
Trainee	47·90%	64·13%	0·007
Competent	55·20%	61·95%	0·532
Expert	65·23%	75·20%	0·097

***Table 2*: Diagnostic accuracy of the EfficientNetV2-based survival benefit prediction (ESBP) system and the clinicians at each expertise level. The first diagnosis represents the accuracy without ESBP assistance, and the second denotes the accuracy with ESBP assistance. The *P* value indicates the statistical significance of the improvement between the two rounds of diagnosis. NA: not applicable.**

These results indicate that the ESBP score could predict survival benefit with a C-index of 0·755 (95% CI: 0·720 −0·791) on the EGFR-TKI training dataset, 0·672 (95% CI: 0·610−0·725) on the EGFR-TKI internal validation dataset, and 0·690 (95% CI: 0·650−0·728) on the EGFR-TKI external test dataset. The proportional hazard assumption evaluation indicated that the ESBP score does not violate the assumption (P = 0·37 and 0·58 on the EGFR-TKI dataset and ICI dataset). The time-dependent ROC curves on the three EGFR-TKI datasets are shown in Supplementary Figure S4.

On the external test dataset used for the reader study, the overall diagnostic accuracy of ESBP was 76·08%, as shown in Table 2. The positive predictive value (PPV) of ESBP, which represents the patients were predicted as good responders truly received a PFS >9·5 months, was 80·40%, 75·40%, and 77·43% on the training, internal validation, and external test datasets, respectively, as shown in Supplementary Table S4.

The results indicated a significant survival benefit difference between the ESBP good responders and first-line chemotherapy NSCLC patients (123 patients with a median PFS of 4·1 months, HR = 0·18, 95% CI = 0·15−0·26, P < 0·0001). Moreover, a statistically significant survival difference was found between the ESBP poor responders and the first-line chemotherapy NSCLC patients (HR = 0·73, 95% CI = 0·50−1·04, P = 0·007). The results are presented in Supplementary Figure S1.

The results as shown in Figure 3 indicate that, when ESBP assisted the trainee, competent, and expert radiology and oncology clinicians, an improvement was achieved in almost all sensitivity, specificity, PPV, and negative predictive value (NPV) indicators. The diagnosis accuracy of the trainee radiologist who had only two years of experience increased from 47·93% to 68·50%, and that of the competent radiologist with five years of experience from 51·03% to 60·86%, reaching a level

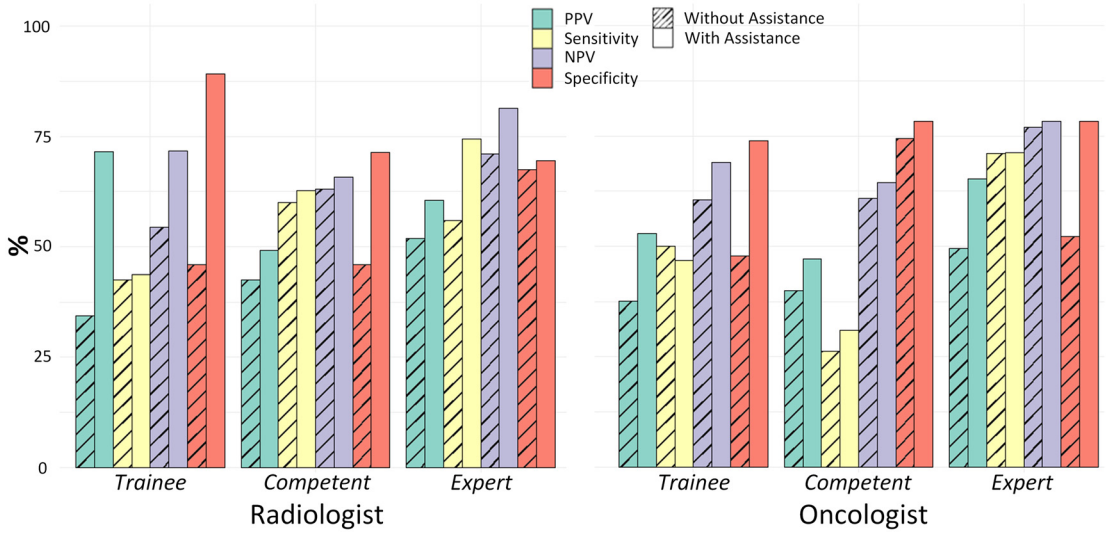

Figure 3. Improvements in the performance of the trainee, competent, and expert clinicians in radiology and oncology with ESBP assistance. Striped bars indicate the result without ESBP assistance. PPV: positive predictive value, NPV: negative predictive value.

similar to that of the expert (65·09%) with more than ten years of radiology experience. Similarly, with ESBP assistance, the diagnostic accuracy of the trainee oncologist improved from 47·90% to 64·13%, and that of the competent oncologist from 55·20% to 61·95%, which is comparable to that of the expert oncologist unassisted by ESBP (65·32%), as shown in Table 2.

The McNemar's test indicated that ESBP significantly improved the performance of the trainee and competent radiologists (both $P < 0·0001$). The three radiologists' first round of diagnoses found no significant difference in the survival benefit between the good responders and poor responders diagnosed by the trainee and competent radiologists ($P = 0·75$ and $P = 0·32$, respectively). When assisted by ESBP, a significant survival difference was found between the poor responders and good responders predicted by the trainee (HR: 2·67, 95% CI: 1·30–5·47, $P < 0·0001$). For the competent radiologist, statistics indicated that the predicted good responders obtained a median PFS of 16·0 months, which was significantly superior to that of the predicted poor responders (median PFS: 10·3 months, HR: 0·26, 95% CI: 0·15–0·46; $P = 0·0031$). When assisted by ESBP, a numerically modest improvement in diagnosis accuracy was observed for the expert radiologist (5·61%). The Kaplan–Meier survival curves of the two rounds of diagnosis by the radiologists are presented in Figure 4.

As shown in Figure 5, no significant survival difference was found between the two subgroups ($P = 0·25$ and $P = 0·78$, respectively) determined by the trainee and competent oncologists in the first-round diagnosis. Assisted by ESBP, the results indicate that a significant diagnosis accuracy improvement of 16·23% was found ($P = 0·007$, McNemar's test) for the trainee oncologist. However, the diagnosis by the competent oncologist showed no significant survival difference in the second round of diagnosis ($P = 0·26$). For the expert oncologist, ESPS assistance significantly improved the survival difference between the predicted good and poor responders in the second round of diagnosis (median PFS: 17·3 months vs. 9·1 months, HR: 0·39, 95% CI: 0·22–0·70, $P < 0·0001$) compared with the first round of diagnosis (median PFS:14·5 months vs. 9·7 months, HR: 0·55, 95% CI: 0·34–0·88, $P = 0·012$).

Without ESBP assistance, the inter-reader agreement (Fleiss' κ) in predicting survival benefit between the radiologist and oncologist with the same level of competence was 0·37 ($P = 0·0004$, trainee), 0·22 ($P = 0·0344$, competent), and 0·42 ($p = 0·0060$, expert), respectively. In addition, when compared to the ESBP model trained by the lung images containing primary tumors, the results indicated that the model based on lung images obtained better diagnostic accuracy than the model based on lung tumor images (90·55% vs. 86·52%) on the training dataset. However, the model based on lung tumor images was more effective on the two test datasets (63·30% vs. 71·73%, and 65·86% vs. 76·08%), as shown in Supplementary Section A3.

To evaluate inter-observer bias using Fleiss' κ, agreements of 0·87 ($P = 0·507$) and 0·89 ($P = 0·651$) were obtained by comparing the scores from the two radiologists' approximate segmentation with the aforementioned scores. Regarding computation, the developed ESBP algorithm was capable of analyzing as many as 118 CT slices per second. After the study, five of the six clinicians stated that they would be willing to use an AI algorithm to triage patients automatically for the prediction of survival benefit for EGFR-TKI therapy.

Discussion

We developed and validated a non-invasive and clinically applicable model to predict the additional survival benefits of both EGFR-TKIs and ICIs in stage IV NSCLC by deep learning analysis of pre-therapy CT images. Additionally, we confirmed that the proposed model was able to triage stage IV NSCLC patients with uncertain survival benefit, and in a reader study, achieved diagnostic accuracy on par with that of experts in radiology and oncology for diagnosing the additional survival benefit of EGFR-TKIs. Moreover, we showed that the deep learning model enabled trainee and competent-level radiologists and oncologists to improve their EGFR-TKI survival benefit diagnosis accuracy to expert level, without additional training.

Studies indicate that in current clinical settings, only 70% of patients with EGFR-positive mutation respond to EGFR-TKI drugs, whereas 18%–45% PD-L1 expressed stage IV NSCLC would respond to ICI therapies.[4,7,8,32] As the largest gene mutation-targeted subpopulation in NSCLC, the EGFR-TKI responders have a median PFS of approximately 9·5 months.[11] Thus, an approach that could accurately predict which patients could gain additional survival benefit through EGFR-TKIs would provide evidence for more precise clinical implementation of EGFR-TKI treatment. However, the assessment of EGFR-TKI benefit for individual stage IV EGFR-mutant NSCLC patients currently lacks sufficient sensitivity to stratify the potential survival benefit for precise clinical decisions. This is further compounded by the low inter-reader agreement on the task ($\kappa < 0·42$ in the three paired clinician groups). The results indicated that the patients in the external test dataset predicted to be good responders of EGFR-TKI by ESBP, obtained a PPV of 77·43% for additional survival benefit evaluation. Therefore, in the clinical guidelines setting, if all stage IV EGFR-mutant NSCLC patients predicted to have a high probability of additional survival benefit were considered for EGFR-TKI therapy, those that would receive additional survival benefit will be significantly increase compared with the current 50% (by median).

The clinical applicability and advancement of existing image-based studies on evaluating the benefit of EGFR-TKIs and ICIs are limited by small sample sizes

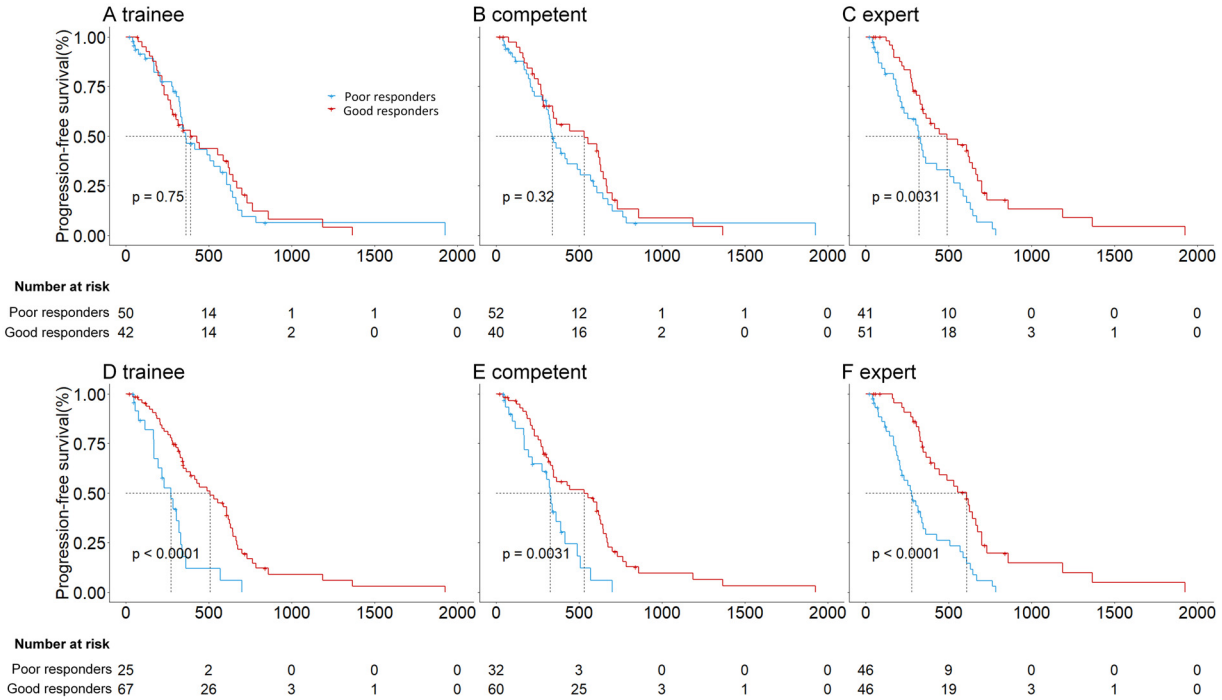

Figure 4. Kaplan–Meier analysis (evaluated by progression-free survival) of the poor and good responders, diagnosed by trainee, competent, and expert radiologists on the external test dataset. A, B, and C represent the results without ESBP assistance, and D, E, and F represent the results with ESBP assistance.

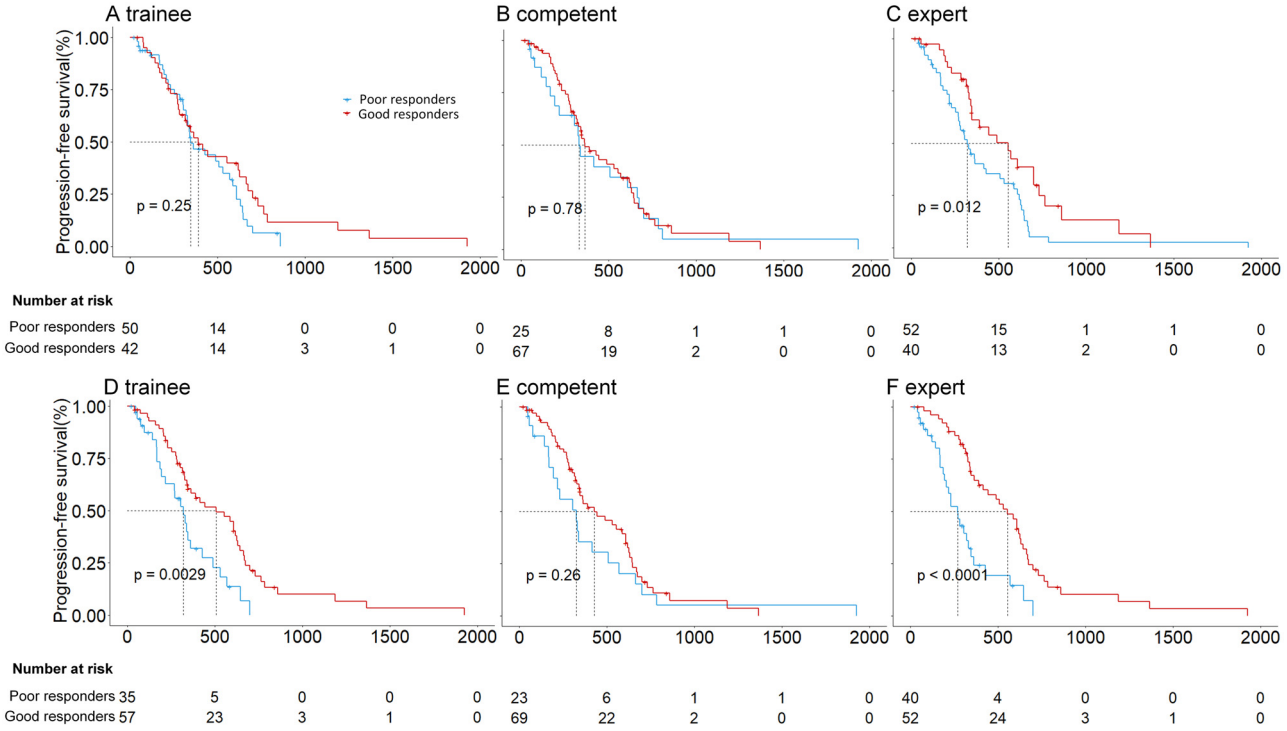

Figure 5. Kaplan–Meier analysis of the two subgroups diagnosed as poor or good responders (evaluated by progression-free survival) by trainee, competent, and expert oncologists on the external test dataset. A, B, and C represent the results without ESBP assistance, and D, E, and F represent the results with ESBP assistance.

and disparate execution models.[33–35] Although previous studies have verified pre-treatment image-based diagnostic tools to assist clinician decision-making of EGFR-TKIs and ICIs administration,[36,37] the results of this study demonstrated that the ESBP score, derived from a clinically accessible CT scan, can evaluate the survival benefit of both EGFR-TKIs and ICIs in stage IV NSCLC patients. Results of the previous study indicated that radiological characteristics such as feature maps, associated with the favorable survival benefit of EGFR-TKIs and ICIs in NSCLC patients, could be decoded by deep learning network.[16] Our study further revealed that the potential commonalities of CT characteristics between the two groups of patients could be decoded via ESBP. In addition, recent studies showed that the activation of EGFR is positively associated with the activation of PD-L1, and the EGFR mutation is correlated with the upregulation of the PD-L1 expression.[38,39] Therefore, further examination of the potential biological processes in combination with radiologic heterogeneity, and examination of the heterogeneous commonalities between the EGFR-mutant NSCLC patients receiving additional survival benefit from EGFR-TKIs, and the PD-L1 expressed patients receiving additional survival benefit from ICI therapy, will contribute to more precise clinical decision.

In addition, the results of this study indicate that ESBP improves both the radiologists' and oncologists' diagnostic accuracy of EGFR-TKI survival benefit in stage IV EGFR-mutant NSCLC patients. By including more extensive datasets from seven participating units, our study investigated the performance of oncologists and radiologists with different degrees of expertise for the task of predicting additional EGFR-TKI survival benefit. When assisted by ESBP, all six clinicians showed improvements in both sensitivity and PPV. In addition, a significant improvement was found in the diagnostic accuracy ($P < 0{\cdot}05$) of trainee and competent radiologists and the trainee oncologist. More notably, the difference in the survival benefit of two subgroups diagnosed by five of the six clinicians showed statistically significant discriminability improvements when assisted by ESBP ($P < 0{\cdot}05$). The results demonstrated that ESBP outperformed previous studies in clarifying the extent to which an AI network can assist radiologists and oncologists to improve survival benefit evaluation, and the advantage of ESBP is to improve the performance of non-expert radiologists and oncologists to a level approximating that of experts without additional training of personnel.

The results predicted that 30·43% of the patients in the EGFR-TKI external test dataset were poor responders (median PFS: 7·2 months), who were unable to receive additional survival benefit through EGFR-TKIs. This finding indicated that more frequent clinical follow-up and monitoring strategies should be adopted for these patients after administering EGFR-TKI therapy. In addition, although a significant difference in PFS was presented between the ESBP predicted poor responders and the advanced NSCLC patients who received first-line chemotherapy ($P = 0{\cdot}007$), the clinical decision-making of first-line chemotherapy vs. targeted therapy still requires caution for the ESBP predicted poor responders. As reported in a previous study, the median PFS of the EGFR-TKI non-responders was only four to five months,[11] which is lower than the median PFS of 7·2 months among the ESBP predicted poor responders in this study.

This study has some limitations. First, as datasets from multiple institutions were involved, it relied on images from multiple scanners and with different kernels and slice thicknesses. In the future, ESBP should be tested for pre-analytical sources of variation, such as scanner manufacturing, reconstruction kernels, and slice thicknesses. The mean of the image-level outputs was used to estimate the patient-level score in this study, and future studies should explore the performance of other fusion approaches. As with other "black-box" neural networks, the methodological understandability of ESBP was not discussed in this study. Efforts to enhance model interpretability could help to increase the suitability of artificial neural network deep learning models. In addition, although the endpoint of PFS has been proved for efficacy evaluation in previous EGFR-TKI prognosis studies,[13,40] other endpoints such as overall survival and safety should be considered during future deep learning model development to facilitate better clinical decision-making. Finally, previous studies have proved the radiologists' manual annotated semantic features for EGFR mutation status prediction,[41,42] combined with the inter-observer evaluation results in this study (Fleiss' $\kappa > 0{\cdot}86$), future work should consider investigating the radiologists' manual annotated semantic features on the survival benefit prediction of EGFR-TKI and ICI therapies.

In conclusion, an artificial intelligence-based system was developed using accessible pre-therapy CT images to accurately predict the survival benefits of EGFR-TKI and ICI therapies for stage IV NSCLC patients. The proposed ESBP assisted non-expert radiologists and oncologists to improve their accuracy of diagnosing EGFR-TKI benefit to expert level. With further optimization and validation on larger, more diverse datasets, the proposed system could offer clinical value as an automated screening tool for selecting patients with stage IV NSCLC for better clinical outcomes.

Contributors

J. Song, K. Deng, P. Gao, and L. Wang conceived the idea for the study with supervision from J. Shi, W. Li, Z. Liu, and J. Tian. K. Deng, J. Shi, W. Li, and Z. Liu conducted the quality assessment. J. Song, K. Deng, P. Gao, and L. Wang drafted the initial draft of the manuscript. All authors contributed to the design, data analysis, and results interpretations. The co-corresponding authors had full access to the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Data sharing statement

The source code, the EfficientNetV2 model, and the sub-dataset for reproducibility experiment are openly available at the following URL link: https://github.com/JD910/ESPS. The data that support the findings of this study are available from the corresponding authors with a signed data access agreement. The raw image and follow-up data are not publicly available because they contain sensitive information that could compromise patient privacy.

Declaration of interests

There are no conflicts of interest to declare.

Acknowledgments

The authors would like to thank all the radiologists and oncologists who participated in this study. This study received funding from the National Natural Science Foundation of China (82001904, 81930053, and 62027901), and Key-Area Research and Development Program of Guangdong Province (2021B0101420005).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.eclinm.2022.101541.

References

1 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249.
2 Fitzmaurice C, Akinyemiju TF, Al Lami FH, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA Oncol*. 2018;4(11):1553–1568.
3 Lin Y, Wang X, Jin H. EGFR-TKI resistance in NSCLC patients: mechanisms and strategies. *Am J Cancer Res*. 2014;4(5):411.
4 Soria JC, Wu YL, Nakagawa K, et al. Gefitinib plus chemotherapy versus placebo plus chemotherapy in EGFR-mutation-positive non-small-cell lung cancer after progression on first-line gefitinib (IMPRESS): a phase 3 randomised trial. *Lancet Oncol*. 2015;16(8):990–998.
5 Soria J-C, Ohe Y, Vansteenkiste J, et al. Osimertinib in untreated EGFR-mutated advanced non–small-cell lung cancer. *N Engl J Med*. 2018;378(2):113–125.
6 Patel SP, Kurzrock R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. *Mol Cancer Ther*. 2015;14(4):847–856.
7 Barnet MB, O'Toole S, Horvath LG, et al. EGFR-co-mutated advanced NSCLC and response to EGFR tyrosine kinase inhibitors. *J Thorac Oncol*. 2017;12(3):585–590.
8 Mazieres J, Drilon A, Lusque A, et al. Immune checkpoint inhibitors for patients with advanced lung cancer and oncogenic driver alterations: results from the IMMUNOTARGET registry. *Ann Oncol*. 2019;30(8):1321–1328.
9 Ettinger DS, Wood DE, Aggarwal C, et al. NCCN guidelines insights: non–small cell lung cancer, version 1.2020: featured updates to the NCCN guidelines. *J Nat l Compr Cancer Netw*. 2019;17(12):1464–1472.
10 Camidge DR, Doebele RC, Kerr KM. Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. *Nat Rev Clin Oncol*. 2019;16(6):341–355.
11 Blumenthal GM, Karuri SW, Zhang H, et al. Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non–small-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. *J Clin Oncol*. 2015;33(9):1008.
12 Xu Y, Hosny A, Zeleznik R, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res*. 2019;25(11):3266–3275.
13 She Y, Jin Z, Wu J, et al. Development and validation of a deep learning model for non–small cell lung cancer survival. *JAMA Netw Open*. 2020;3(6):e205842-e.
14 Baikejiang R, Giltnane J, Fuentes E, Kozlowski C. *Abstract PO-088: A Deep-Learning Based Approach to Assess Heterogeneity of Histologies in Non-Small Cell Lung Cancer*. AACR; 2020.
15 Lai Y-H, Chen W-N, Hsu T-C, Lin C, Tsao Y, Wu S. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Sci Rep*. 2020;10(1):1–11.
16 Mu W, Jiang L, Zhang J, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. *Nat Commun*. 2020;11(1):5228.
17 Hou R, Li X, Xiong J, et al. Predicting tyrosine kinase inhibitor treatment response in stage IV lung adenocarcinoma patients with EGFR mutation using model-based deep transfer learning. *Front Oncol*. 2021;11:679764.
18 Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning; 2019: PMLR*. 2019. p. 6105–6114.
19 Marques G, Agarwal D, de la Torre Díez I. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Appl Soft Comput*. 2020;96:106691.
20 Tan M, Le Q. Efficientnetv2: smaller models and faster training. *International Conference on Machine Learning; 2021: PMLR*. 2021. p. 10096–10106.
21 Hirai K, Kuwahara T, Furukawa K, et al. Artificial intelligence-based diagnosis of upper gastrointestinal subepithelial lesions on endoscopic ultrasonography images. *Gastric Cancer*. 2022;25(2):382–391.
22 Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–247.
23 Mauguen A, Pignon J-P, Burdett S, et al. Surrogate endpoints for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer: a re-analysis of meta-analyses of individual patients' data. *Lancet Oncol*. 2013;14(7):619–626.
24 Hotta K, Fujiwara Y, Matsuo K, et al. Time to progression as a surrogate marker for overall survival in patients with advanced non-small cell lung cancer. *J Thorac Oncol*. 2009;4(3):311–317.
25 Dercle L, Fronheiser M, Lu L, et al. Identification of non–small cell lung cancer sensitive to systemic cancer therapies using radiomics. *Clin Cancer Res*. 2020;26(9):2151–2162.
26 Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res*. 2004;10(21):7252–7259.
27 Silva F, Pereira T, Morgado J, et al. EGFR assessment in lung cancer CT images: analysis of local and holistic regions of interest using deep unsupervised transfer learning. *IEEE Access*. 2021;9:58667–58676.
28 Song J, Yang C, Fan L, et al. Lung lesion extraction using a toboggan based growing automatic segmentation approach. *IEEE Trans Med Imaging*. 2015;35(1):337–353.
29 Lakatos E. Designing complex group sequential survival trials. *Stat Med*. 2002;21(14):1969–1989.
30 Falotico R, Quatto P. Fleiss' kappa statistic without paradoxes. *Qual Quant*. 2015;49(2):463–470.
31 Adedokun OA, Burgess WD. Analysis of paired dichotomous data: a gentle introduction to the McNemar test in SPSS. *J MultiDiscip Eval*. 2012;8(17):125–131.
32 Arrieta O, Barron F, Padilla MS, et al. Effect of metformin plus tyrosine kinase inhibitors compared with tyrosine kinase inhibitors alone in patients with epidermal growth factor receptor-mutated lung adenocarcinoma: a phase 2 randomized clinical trial. *JAMA Oncol*. 2019;5(11):e192553.
33 Cook GJ, O'Brien ME, Siddique M, et al. Non-small cell lung cancer treated with erlotinib: heterogeneity of (18)F-FDG uptake at PET-association with treatment response and prognosis. *Radiology*. 2015;276(3):883–893.
34 Park S, Ha S, Lee SH, et al. Intratumoral heterogeneity characterized by pretreatment PET in non-small cell lung cancer patients predicts progression-free survival on EGFR tyrosine kinase inhibitor. *PLoS One*. 2018;13(1):e0189766.

35 Takeda M, Okamoto I, Nakagawa K. Survival outcome assessed according to tumor response and shrinkage pattern in patients with EGFR mutation-positive non-small-cell lung cancer treated with gefitinib or erlotinib. *J Thorac Oncol*. 2014;9(2):200–204.

36 Song J, Wang L, Ng NN, et al. development and validation of a machine learning model to explore tyrosine kinase inhibitor response in patients with stage IV EGFR variant-positive non-small cell lung cancer. *JAMA Netw Open*. 2020;3(12):e2030442.

37 Liu Y, Kim J, Balagurunathan Y, et al. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. *Clin Lung Cancer*. 2016;17(5):441-8. e6.

38 Akbay EA, Koyama S, Carretero J, et al. Activation of the PD-1 pathway contributes to immune escape in EGFR-driven lung tumors. *Cancer Discov*. 2013;3(12):1355–1363.

39 Azuma K, Ota K, Kawahara A, et al. Association of PD-L1 overexpression with activating EGFR mutations in surgically resected nonsmall-cell lung cancer. *Ann Oncol*. 2014;25(10):1935–1940.

40 Park K, Tan E-H, O'Byrne K, et al. Afatinib versus gefitinib as first-line treatment of patients with EGFR mutation-positive non-small-cell lung cancer (LUX-Lung 7): a phase 2B, open-label, randomised controlled trial. *Lancet Oncol*. 2016;17(5):577–589.

41 Gevaert O, Echegaray S, Khuong A, et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Sci Rep*. 2017;7 (1):1–8.

42 Zhou M, Leung A, Echegaray S, et al. Non—small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology*. 2018;286(1):307–315.