Machine learning analysis of pathological images to predict 1-year progression-free survival of immunotherapy in patients with small-cell lung cancer

Ryota Shibaki,¹ Daichi Fujimoto,¹ Tsukasa Nozawa,² Akira Sano,² Yuka Kitamura,³ Junya Fukuoka,³ Yuki Sato,⁴ Takashi Kijima,⁵ Hirotaka Matsumoto,⁶ Toshihide Yokoyama,⁷ Satoru Miura,⁸ Akito Hata,⁹ Motohiro Tamiya,¹⁰ Yoshihiko Taniguchi,¹¹ Jun Sugisaka,¹² Naoki Furuya,¹³ Hisashi Tanaka,¹⁴ Nobuyuki Yamamoto,¹,¹⁵ Yasuhiro Koh,¹,¹⁵ Hiroaki Akamatsu¹

ABSTRACT

Background: In small-cell lung cancer (SCLC), the tumor immune microenvironment (TIME) could be a promising biomarker for immunotherapy, but objectively evaluating TIME remains challenging. Hence, we aimed to develop a predictive biomarker of immunotherapy efficacy through a machine learning analysis of the TIME.

Methods: We conducted a biomarker analysis in a prospective study of patients with extensive-stage SCLC who received chemoimmunotherapy as the first-line treatment. We trained a model to predict 1-year progression-free survival (PFS) using pathological images (H&E, programmed cell death-ligand 1 (PD-L1), and double immunohistochemical assay (cluster of differentiation 8 (CD8) and forkhead box P3 (FoxP3))) and patient information. The primary outcome was the mean area under the curve (AUC) of machine learning models in predicting the 1-year PFS.

Results: We analyzed 100,544 patches of pathological images from 78 patients. The mean AUC values of patient pathological images, and combined models were 0.789 (range 0.571–0.982), 0.782 (range 0.750–0.911), and 0.868 (range 0.786–0.929), respectively. The PFS was longer in the high efficacy group than in the low efficacy group in all three models (patient information model, HR 0.468, 95% CI 0.868 (range 0.786–0.929); pathological image model, HR 0.334, 95% CI 0.171 to 0.628; combined model, HR 0.353, 95% CI 0.195 to 0.637). The machine learning analysis of the TIME had better accuracy than the human count evaluations (AUC of human count, CD8-positive lymphocyte: 0.681, FoxP3-positive lymphocytes: 0.625, PD-L1 score: 0.567).

Conclusions: The spatial analysis of the TIME using machine learning predicted the immunotherapy efficacy in patients with SCLC, thus supporting its role as an immunotherapy biomarker.

INTRODUCTION

Lung cancer is one of the primary causes of cancer-related deaths.¹ Small-cell lung cancer (SCLC), which accounts for 15% of all lung cancer cases, is characterized by rapid proliferation, early development of widespread metastases, and poor survival rates.² ³ Combining immunotherapy with chemotherapy has durable clinical activity and manageable safety.⁴ ⁵ Hence, chemoimmunotherapy is recommended as first-line treatment for extensive-stage (ES) SCLC.⁶

Notably, long-term efficacy is a hallmark feature of immunotherapy; however, a 1-year progression-free survival (PFS) has only been observed in 10%–20% of patients with SCLC.⁴ ⁵ Thus, the development of biomarkers for predicting the efficacy of
immunotherapy is required as no definitive biomarkers have been established for SCLC. Recently, the concept of the tumor immune microenvironment (TIME), where the activation and suppression of lymphocytes and their localization with the tumor take place, has attracted attention as a biomarker for immunootherapy. The localization of immune cells (cluster of differentiation 8 (CD8)-positive cells and forkhead box P3 (FoxP3)-positive cells) reflects the host antitumor immunity status.

Several important clinical issues should be considered when evaluating the TIME. The spatial analysis of the relationship between inflammatory cells and the tumor is important for the accurate evaluation of the TIME. However, due to the complexity of interactions between various inflammatory cells and tumor cells, it is difficult to determine which interactions are important for enhancing the immunotherapeutic efficacy. Furthermore, objectively evaluating such spatial analysis remains challenging. Recently, machine learning has attracted significant attention for standardization, reproducibility, and improving accuracy.

Therefore, to develop an objective biomarker for assessing the efficacy of immuno-therapy using the TIME, the precise and objective spatial analysis of the complex positioning of inflammatory and tumor cells should be performed using machine learning. Another issue in examining the usefulness of the TIME as a biomarker in clinical practice is that the patient’s background affects the TIME and the efficacy of immunotherapy; thus, it is necessary to determine whether the TIME status is a more accurate predictor of immunotherapeutic efficacy than patient information through the conduct of prospective cohort studies that involve the collection of high-quality patient information. Therefore, we aimed to investigate the feasibility of developing biomarkers for determining the efficacy of immuno-therapy in prospective cohorts of ES-SCLC patients through a machine learning analysis of the TIME and patient information.

**METHODS**

**Patients**

This 32-centered, prospective, and hospital-based cohort study was a biomarker analysis of the APOLLO study, using patients with ES-SCLC who received carboplatin and etoposide with atezolizumab as first-line treatment in Japan between September 2019 and September 2020. In addition to the eligibility criteria in the APOLLO study, patients in whom adequate biopsy specimens were obtained before initiating the first-line treatment were included. We only used biopsy samples from the lungs to accurately evaluate the TIME status. The cut-off date for data collection was September 2021.

The study was registered in the University Medical Information Network Clinical Trials Registry (UMIN 000039600).

**Processing of pathological images**

The tissue samples were fixed in 10% formalin and embedded in paraffin. Immunohistochemical analysis was performed on 4 μm thick, formalin-fixed, paraffin-embedded tissue sections. Each section was deparaffinized in xylene, rehydrated, and incubated in 0.05% H2O2 in 95% methanol for 10 min. All samples were stained with H&E, programmed cell death-ligand 1 (PD-L1, Dako PD-L1 immunochemistry (IHC) 28-8 pharmDx assay), and chromogen-based double immunohistochemical assay (CD8 (clone: C8/144B) and FoxP3 (clone: 236A/E7)) based on the results of recent studies.

For double staining using primary antibodies from the same species, inactivation of the primary antibody and linked enzymes used was performed. Slides after the first staining with AEC (3-amino-9-ethylcarbazole) were washed in phosphate-buffered saline for 5 min, autoclaved at 110°C for 10 min in the epitope retrieval solution for the second antigen, and the second staining with DAB (diaminobenzidine) was carried out. After staining, sections were counterstained with Myer’s hematoxylin, dehydrated with xylene, and permanently mounted. Representative images of H&E, PD-L1, and double IHC are provided in online supplemental figure 1. All slides were scanned using a whole-slide imaging scanner. The assigned pathologists (YKitamura and JF) manually annotated the tumor area in all samples on the H&E staining images to assess the relationship between each inflammatory cell and the tumor. In CD8 and Foxp3 multiplex staining slides, the number of each type of infiltrating immune cell was counted in up to four 1 mm² fields, and the mean tumor infiltration lymphocyte (TIL) count was recorded. The tumor cell PD-L1 expression was defined as the percentage of viable tumor cells with partial or complete membrane staining in at least 100 viable tumor cells. The Combined Positive Score (CPS) was generated by scoring the PD-L1-stained slides using the CPS algorithm; it was calculated by dividing the number of PD-L1-positive tumor cells, lymphocytes, and macrophages by the total number of viable tumor cells and multiplying the results by 100.

**Development of the predictive model**

We aimed to develop a predictive model for 1-year PFS using a machine learning algorithm. In previous phase III trials, the 1-year PFS rate was 5.3%–5.4% in the chemotherapy group, compared with 12.6%–17.9% in the chemoimmunotherapy group. Therefore, the implication is that most of the patients who achieved 1-year PFS did so due to immune checkpoint immunotherapy. A two-class classifier that predicts the 1-year PFS was trained using the whole-slide pathological images of H&E, PD-L1 expression, double IHC (CD8 and FoxP3), and patient information (figure 1). The predictive model was developed in three steps: “preprocessing of patch image dataset (figure 1A,B),” creation of a “patch-based pipeline (figure 1C,D),” and completion of “per-patient pipeline (figure 1E,F).” During the model development,
our network was trained using a fivefold cross-validation approach that involved randomly splitting the full set of our patch image dataset into five balanced subsets (figure 1G). The balanced subsets do not refer to a simple random split. Instead, we divided all patients into five groups so that the ratio of 1-year PFS responders and non-responders was equal. Then, patches were generated from each patient and used in the analyses, so that there was no leakage. After assessing the predictive model, the coordinates for the optimal threshold were identified on the receiver operating characteristic (ROC) curves from each fold. The Youden index from the optimal point on the ROC curve (distance from 0 to 1) was used to determine the optimal threshold values, which were then set as cut-off points for stratifying patients into high or low efficacy groups.

Preprocessing of the patch image dataset

For preprocessing, the three types of whole-slide images were processed. The tumor area annotation was transferred from the H&E staining images to the other types of images. Since each whole-slide image is not in the same position, it is impossible to use the same annotation directly. In order to avoid this problem, we applied the following two processes for the image alignment process.

First, the same tumor region between each image was manually annotated. Second, a non-rigid image registration framework was constructed using the “Insight Segmentation and Registration Toolkit”24 (figure 1A). Next, we eliminated the background by setting a threshold for white color. In the final step, we divided the inside of the annotation into 128-pixel square images without overlap (figure 1B). By going through the process up to this point, N patch images of 128 pixels in height and width are generated for each patient. N represents the total number of patches and has a different number for each patient.

Patch-based pipeline

In creating the patch-based pipeline step, the confidence score for the 1-year PFS for each patch was predicted. The creation of a patch-based pipeline involved two phases: “feature extraction module” (figure 1C) and “patch-based classifier” (figure 1D). We used EfficientNet,25 a widely used architecture for image recognition tasks, as a feature extraction module. Specifically, a pretrained EfficientNet-B0 model that was trained on the ImageNet dataset was used.26 Since we wanted to use image features that were learned on a large variety of images directly, retraining and fine-tuning of the EfficientNet model were

Figure 1 Overview of the predictive model for the achievement of 1-year progression-free survival (PFS). (A) HE, PD-L1, and double immunohistochemistry (CD8 and FoxP3) analyses were performed in all samples. The green contour represents the annotation of the tumor area. (B) Patches from the same position within the annotated area. The patch size had a height of 128 pixels and a width of a patch size. (C) A pretrained EfficientNet-B0 model was used. (D) A patch-based pipeline predicted the confidence score for 1-year PFS for each patch. (E) Patient information combined the histogram confidence score calculated from the patch-based pipeline. (F) The patient-based classifier, which uses LightGBM, predicts the confidence score for 1-year PFS. (G) The overall patient data were divided into five randomly chosen folds of equal size. One-fold was used to validate the model trained using the remaining folds. (H) The overall patients (including the training and validation cohorts) were classified into high and low efficacy groups using the predictive model from each fold. BatchNorm, batch normalization layer; CD8, cluster of differentiation 8; Conv, convolution layer; FC, fully connected layer; LeakyReLU, leaky rectified linear unit layer; MBConv, Mobile Inverted Bottleneck Convolution Layer; PD-L1, programmed cell death ligand 1; Train, training cohort; Val, validation cohort.
not performed. After completing the “feature extraction module,” the features consisting of 1280 dimensions were obtained for each patch. To calculate the final score based on the features, a patch-based classifier consisting of three fully connected layers was constructed. First, three features were combined as the input of the patch-based classifier. Then, 3840 dimensional features were used as the input, and a two-class classification was performed. The estimated class and target were determined using the sum of the cross-entropy loss and were similar for all patches. The PyTorch framework was used to train the network with graphical processing units. The drop-out rates were set to 25% for the first layer and 50% for the second layer to reduce overfitting. The Adam optimizer was used for the optimization of stochastic gradient descent with a batch size of 48.

Per-patient pipeline

In this step, the histogram features calculated from the results of the patch-based prediction were generated for each patient, and the patient-based prediction was performed by concatenating the histogram features and patient information (figure 1E). The 1-year achievement probability per patch from the patch-based classifier was used for the histogram. A histogram confidence score of 20 bins was obtained. Since the number of patches for each patient differed, each histogram was normalized to the sum of 1. Thus, a histogram confidence score 0%–5% showed the proportion of the patches that predicted 0%–5% probability among all patches for a patient. Then, the per-patient image features were combined with the corresponding patient information. The following 21 pieces of information were used: sex, age, Eastern Cooperative Oncology Group performance status score, smoking status, TNM classification, history of autoimmune disease, systemic immunosuppressive therapy status, pre-existing interstitial lung disease, history of malignancy other than SCLC within the last 5 years, history of thoracic radiotherapy, brain metastasis, liver metastasis, bone metastasis, malignant pleural effusion and pericardial effusion, adrenal metastasis, neutrophil count, platelet count, hemoglobin level, serum bilirubin level, serum aspartate transaminase and alanine transaminase levels, and serum creatinine level. This 43-dimensional feature was defined as the per-patient information. Finally, a patient-based classifier was prepared to predict the confidence score for the 1-year PFS of each patient (figure 1F). The Gradient Boosting Decision Tree algorithm was used as a classifier, while the LightGBM library was used to train the classifier. LightGBM was also used in the patient information model, the pathological image model, and the combined model; these were performed in 23, 20, and 43 dimensions, respectively.

Statistical analysis

The primary outcome was the accuracy of machine learning models in predicting the 1-year PFS. The area under the ROC curve was calculated to evaluate the accuracy of the machine learning model. In addition, the area under the ROC curve was calculated to evaluate the predictive accuracy of CD8 count, FoxP3 count, and PD-L1 CPS. The Kaplan-Meier method was used to estimate the PFS, and the groups were compared using the Cox proportional hazard model. Results were expressed as HRs and 95% CIs. An area under the curve (AUC) of ≥0.7 was used to denote a good predictive value, and a two-sided p<0.05 was considered significant.

RESULTS

Patients’ characteristics

We analyzed 78 patients from a prospective cohort of 207 patients with ES-SCLC who received chemoimmunotherapy as first-line treatment. The overall study result was already published elsewhere. The patient selection process is shown in online supplemental figure 2, while the baseline clinical characteristics are shown in table 1. The patients’ median age was 78 years. Thirty patients (38%) were aged ≥75 years. A total of 67 patients (86%) had a performance status score of 0 or 1 when first-line therapy was initiated. The median follow-up period was 394 days (IQR: 200–484), and the minimum follow-up period for all patients was 365 days. The overall objective response rate, median PFS, and overall survival (OS) of all patients were 72.4% (55/78), 145 days, and 394 days, respectively. The 1-year PFS rate was 10.3% (8/78).

Table 1 Baseline clinical characteristics of the patients

<table>
<thead>
<tr>
<th>All n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>≥75 years</td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>ECOG PS</td>
</tr>
<tr>
<td>0–1</td>
</tr>
<tr>
<td>≥2</td>
</tr>
<tr>
<td>Smoking status</td>
</tr>
<tr>
<td>Never-smoker</td>
</tr>
<tr>
<td>Current or former smoker</td>
</tr>
<tr>
<td>Liver metastasis</td>
</tr>
<tr>
<td>Bone metastasis</td>
</tr>
<tr>
<td>Adrenal metastasis</td>
</tr>
<tr>
<td>Brain metastasis</td>
</tr>
<tr>
<td>Malignant pleural effusion and pericardial effusion</td>
</tr>
<tr>
<td>Treatment with systemic immunosuppressive therapy</td>
</tr>
<tr>
<td>ECOG PS, Eastern Cooperative Oncology Group performance status</td>
</tr>
</tbody>
</table>
Accuracy of predictive models for treatment efficacy using machine learning analysis

Among analyzed patients, 8 patients achieved 1-year PFS, whereas 70 patients did not. We obtained 100,544 patches of pathological images from all patients; of these, 9878 patches were from patients that achieved 1-year PFS, and 90,666 patches were from those that did not achieve 1-year PFS. Using a fivefold cross-validation approach, the predictive models showed the following mean AUC values: 0.789 (95% CI 0.671 to 0.922) for the patient information model, 0.782 (95% CI 0.735 to 0.844) for the pathological image model, and 0.868 (95% CI 0.818 to 0.910) for the combined model. The ranges of AUC values across all folds were 0.571–0.982, 0.750–0.911, and 0.786–0.929 in the patient information model, pathological image model, and combined model, respectively. The ROC curves are shown in figure 2. In addition, the predictive models showed the following mean c-indices: 0.731 (95% CI 0.656 to 0.807) for the patient information model, 0.754 (95% CI 0.725 to 0.783) for the pathological image model, and 0.820 (95% CI 0.749 to 0.892) for the combined model. To confirm the accuracy of this analysis, we excluded the three patients that were censored within 1 year and retrained the model a total of 75 patients. The predictive models showed the following mean AUC values: 0.754 (95% CI 0.695 to 0.812) for the patient information model, 0.786 (95% CI 0.747 to 0.824) for the pathological image model, and 0.818 (95% CI 0.745 to 0.855) for the combined model. The ROC curves are shown in online supplemental figure 4.

Predictive model stratifying patients treated with immunotherapy by PFS

The patients were classified into high and low efficacy groups using the predictive model to examine its predictive accuracy for PFS. Since five models were created using a fivefold cross-validation approach, the PFS of the high and low efficacy groups was estimated for each fold. The Kaplan-Meier curves for PFS are shown in figure 3 for each median accuracy model and in online supplemental figure 3 for all folds. The PFS was longer in the high efficacy group than that in the low efficacy group when determined using the patient information model (HRs (95% CIs) 0.714 (0.425 to 1.202), 0.676 (0.422 to 1.082), 0.468 (0.287 to 0.762), 0.507 (0.281 to 0.913), and 0.648 (0.400 to 1.049) for each fold, respectively), pathological image model (HRs (95% CIs) 0.089 (0.026 to 0.303), 0.125 (0.044 to 0.357), 0.334 (0.177 to 0.628), 0.411 (0.236 to 0.714), and 0.524 (0.312 to 0.881) for each fold, respectively), and combined model (HRs (95% CIs) 0.152 (0.062 to 0.370), 0.266 (0.141 to 0.500), 0.353 (0.195 to 0.637), 0.411 (0.236 to 0.714), and 0.420 (0.240 to 0.732) for each fold, respectively). Further, the Kaplan-Meier curves for PFS of the model excluding the patients censored within 1 year are shown in online supplemental figure 5.

Interpretative importance score from the trained model

The relative importance of a specific variable for determining the 1-year PFS achievement was calculated using each predictive model; the importance of each model is shown in table 2. Since we used LightGBM to train final classifiers, the importance of each element can be calculated by focusing on how many times each element was used in a branch during the learning process.32 The patient information model demonstrated that age was the most important factor, followed by the presence of malignant pleural effusion and pericardial effusion, smoking status, presence of bone metastasis, and presence of liver metastasis. In the predictive model based on the pathological images, the histogram confidence score consisting.
of 20 bins was calculated based on the results of patch-based prediction for each patient. The histogram confidence score reflects the percentage of patches predicting the probability of the 1-year PFS achievement in patch-based prediction. The pathological image model demonstrated that the histogram confidence score of 0%–5% had the highest importance, followed by histogram confidence scores of 95%–100%, 15%–20%, 5%–10%, and 10%–15%. The combined model demonstrated that a histogram confidence score of 0%–5% had the highest importance, followed by the histogram confidence scores of 5%–10%, 10%–15%, 95%–100%, and 15%–20%. The pathological image features obtained scores with higher importance in the combined model than the scores of the patient information features.

Accuracy in predicting the achievement of 1-year PFS based on the manual cell count

To examine the usefulness of machine learning-based evaluation, the predictive value of manually analyzed pathological images was determined. The densities (medians (ranges)) of CD8-positive TILs and FoxP3-positive TILs were 18 (0–668)/mm² and 32 (0–541)/mm², respectively. The median (range) of PD-L1 CPS was 0 (0–100). The densities of CD8-positive TILs and FoxP3-positive TILs were comparable between patients who achieved and those who did not achieve the 1-year PFS (CD8-positive TILs: median, 52/mm² vs 23/mm², p=0.286; FoxP3-positive TILs: median, 42/mm² vs 30/mm², p=0.396). The PD-L1 CPS was also comparable between patients who achieved and those who did not achieve the 1-year PFS (median, 0 vs 0; p=0.325). For the 1-year PFS, the CD8-positive TILs, FoxP3-positive TILs, and PD-L1 CPS demonstrated the AUC values of 0.681, 0.626, and 0.567, respectively. The ROC curves are shown in figure 4.

DISCUSSION

Here, we used prospective cohorts and investigated the predictive value of a machine learning model that used pathological images and patient information for determining the efficacy of immunotherapy in patients with ES-SCLC who received chemoimmunotherapy as first-line treatment. The pathological image model was found to have a good performance in determining the efficacy of immunotherapy. In the combined model using patient information and pathological images, the importance of the pathological images in predicting the achievement of 1-year PFS was higher than that of patient information. To the best of our knowledge, this is the first study to show that a classification of the TIME using machine learning could be a biomarker for immunotherapy efficacy in SCLC.

The primary objective of our study was to identify a potential biomarker of immunotherapy efficacy in SCLC, where biomarkers have not been established. Thus far, PD-L1 expression, tumor mutation burden, and microsatellite instability have been approved by the Food and Drug Administration as companion biomarkers.
of immunotherapy response in other tumor types. The reported AUC range for these companion biomarkers was 0.685–0.755, which denotes their utility in clinical practice. In our pathological image model, the mean AUC for the treatment efficacy was 0.782, which was comparable to that of the approved companion biomarkers. Furthermore, the AUC with the lowest accuracy among the fivefold was 0.750. This result may indicate that the pathological image model had good generalization performance as each model displayed high accuracy regardless of the randomness of the patient population. Therefore, the prediction using our pathological image model has sufficiently high predictive performance compared with the approved companion biomarkers of other tumor types and can be a biomarker of immunotherapy efficacy in SCLC.

This study showed that the pathological images were more important for predicting the efficacy of immunotherapy than the patient information. In several tumor types, the efficacy of immunotherapy differs depending on patient factors such as age, smoking status, metastasis site, and performance status. Although these factors were important in the patient information model, “age” was the only patient information-related factor that was among the top 10 most important factors in the combined model, while the rest were pathological image factors. Our findings have shown that the TIME may play a more important role in predicting the antitumor efficacy of immunotherapy than the patient background. Hence, the TIME status should be considered when developing future biomarkers for determining immunotherapy efficacy.

Machine learning-based analysis outperforms manual TIL assessment when predicting the efficacy of immunotherapy, although the underlying reason or mechanism remains unknown. However, the spatial analysis of the relationship between each inflammatory cell and tumor is important for making accurate predictions. In previous studies, the spatial evaluation of TILs based on nuclear segmentation and deep convolutional neural networks from pathological images were used to predict immunotherapy efficacy. However, these studies were limited by difficulties in performing a spatial analysis that included the tumor area and the role of lymphocytes. In our study, tumor area annotation and lymphocyte IHC were used to evaluate the tumor and lymphocyte localization. Lymphocyte localization (stromal or intratumoral) affected the lymphocyte activity when examining the antitumor immune status. Thus, a spatial analysis of the relationship between each inflammatory cell and the tumor is important for evaluating the TIME status, which can predict the immunotherapy efficacy.

When evaluating the TIME, machine learning analysis has better objectivity, which is essential for enhancing the reliability of a biomarker in clinical practice. However, a manual spatial analysis performed by a human has a significant interobserver variability that can negatively impact the findings of studies and clinical trials. Using machine learning, we were able to conduct an objective spatial analysis that humans cannot achieve. Based on these perspectives, machine learning will be important for analyzing complex pathological images for use as a biomarker in clinical practice.

Our predictive biomarker was developed using only three pathological slides without the need for any special equipment or techniques. Although RNA analysis has been suggested for classifying the immune status of SCLC patients, it requires large amounts of specimens. In general, patients with advanced-stage lung cancer are often diagnosed through a bronchoscopic examination, but many of them are unable to undergo biomarker testing due to the inadequate tissue volume obtained. Therefore, testing using a small number of specimens or a small specimen size is critical in clinical practice. Moreover, H&E staining and IHC do not require special equipment and techniques. Thus, our method is relatively simple for identifying reliable biomarkers in routine clinical practice.

Our study has some limitations. First, the amount of available data was limited. This created a ceiling for optimizing the machine learning analysis because the weights and biases are trained. However, a certain level of classification performance acceptable for clinical application can be achieved using only 100 cases, and the number of cases required for model training is declining. Second, although a fivefold cross-validation was conducted, external independent cohort validation was not performed. Therefore, our predictive model may be overfitting and may overestimate the model accuracy. In fact, it was not possible to evaluate all folds with a common threshold, such as one calculated from the Youden index of the average ROC. However, all of the models showed the same tendency in terms of importance that developed final model. Although further validation is needed with more data in the future, we believe that our models were able to extract features and model for key of biomarker. Finally, because our predictive model was developed by deep learning and validated on the results of H&E staining, the PD-L1 expression determined via IHC, and the CD8_FoxP3 IHC pathological images in parallel, the specific contribution of each pathological image to the prediction model and the interpretation of the TIME remains unclear.

In conclusion, a machine learning analysis of TIME using pathological images can predict the efficacy of immunotherapy in SCLC, which has no established biomarker. An objective spatial analysis is critical for evaluating the TIME status and predicting the efficacy of immunotherapy. Evaluating the TIME using machine learning analysis may help determine the efficacy of immunotherapy.
Acknowledgements We would like to thank the patients, clinical staff, data manager of Osaka Metropolitan University, and other support staff of Wakayama Medical University.

Contributors RS: conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing—original draft, and writing—review and editing. DF: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, writing—original draft, writing—review and editing, and the guarantor. TN and AS: formal analysis (machine learning analysis), methodology, writing—original draft, and writing—review and editing. YKitamura and JF: formal analysis (pathological analysis), methodology, writing—original draft, and writing—review and editing. YS, TK, HM, TY, SM, AH, MT, YT, JS, NF, HT, NY and YOkh. HA: conceptualization, data curation, investigation, writing—original draft, and writing—review and editing.

Funding This work was supported by Chugai Pharmaceutical.

Competing interests The authors declare the following financial interests/personal relationships, which may be considered potential competing interests. DF has received personal fees from AstraZeneca KK, Boehringer Ingelheim Japan; Ono Pharmaceutical; Bristol-Myers Squibb; Taiho Pharmaceutical; Chugai Pharmaceutical; MSD KK; Eli Lilly Japan KK; Kyowa Kirin; and Novartis Pharma KK, outside of the submitted work. YKitamura is the chief executive officer of N Lab Co., Ltd and its stockholder. JF is an advisor of N Lab Co., Ltd. YS has received personal fees from AstraZeneca, Chugai Pharmaceutical, MSD, Ono Pharmaceutical, Novartis, Pfizer, Taiho Pharmaceutical, nippon Kayaku, Bristol-Myers Squibb, Eli Lilly, Takeda, and Kyowa Kirin, outside of the submitted work. TK has received personal fees and scholarship donations from Chugai Pharmaceutical, outside of the submitted work. HM has received personal fees from Chugai Pharmaceutical, outside of the submitted work. TY has received personal fees from Chugai Pharmaceutical, outside of the submitted work. SM has received personal fees from Chugai Pharmaceutical, Taiho Pharmaceutical, Ono Pharmaceutical, AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Boehringer-Ingehelm Japan, and Takeda Pharmaceutical, outside of the submitted work. AH has received personal fees from Chugai outside of this study, outside of the submitted work. YT has received personal fees from Chugai outside of this study, outside of the submitted work. HT has received personal fees from Chugai Pharmaceutical, Ono Pharmaceutical, AstraZeneca, Bristol-Myers Squibb, Boehringer-Ingehelm Japan, and Pfizer Japan, outside of the submitted work. NY has received research funds from Chugai Pharmaceutical, outside of the submitted work.

Patient consent for publication Consent obtained directly from patient(s).

Ethics approval The Wakayama Medical University institutional review board (No. 2805) approved the protocol. The study protocol and all amendments were approved by the independent institutional review board of each participating institution. The study was conducted in accordance with the principles of the Declaration of Helsinki. All patients provided written informed consent prior to study enrolment.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The corresponding author declares that he had full access to the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. The data supporting the findings of this study are available from the corresponding author on reasonable request. The data are not publicly available due to privacy and ethical restrictions.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See http://creativecommons.org/licenses/by-nc/4.0/

ORCID ids Daichi Fujimoto http://orcid.org/0009-0002-6094-4858 Takashi Kijima http://orcid.org/0000-0003-4249-4021 Satoru Milura http://orcid.org/0000-0001-7868-6661

REFERENCES

Small cell lung cancer treated with immune checkpoint inhibitors—
Dall'Olio FG, Maggio I, Massucci M, et al. Outcomes of chemotherapy and survival trends in advanced non-
Grilly-Olson JE, Hayes DN, Moore DT, et al. Validation of interobserver agreement in lung cancer assessment: hematoxylin-
**Supplementary Figure 1** Pathological images. 

- **a** Hematoxylin and eosin.
- **b** Programmed cell death-ligand 1 (Dako PD-L1 IHC 28-8 pharmDx assay).
- **c, d** Chromogen-based double immunohistochemical assay [CD8 (clone: C8/144B) and FoxP3 (clone: 236A/E7), black arrow is a double-positive cell]
APOLLO study
(n=207)

Pathology images were available
(n=126)

Excluded (n=81)
78 had no pathological specimens
3 had no tumor annotation

Excluded (n=48)
41 tissue samples were not collected from the lungs.
7 were not processed for data.

Analyzed (n=78)

Supplementary Figure 2  Flowchart of the participant selection process
Supplementary Figure 3 Kaplan–Meier curves for the progression-free survival of the high and low efficacy groups predicted using our predictive models for all folds. a Patient information model. b Pathological image model. c Combined model
Performance of the patient-based classifier without the censored patients within one-year was evaluated using the area under the receiver operating characteristic curve value for one-year progression-free survival.

**Supplementary Figure 4**

- **a. Patient information model**
  - ROC Fold 0 (AUC=0.661)
  - ROC Fold 1 (AUC=0.839)
  - ROC Fold 2 (AUC=0.714)
  - ROC Fold 3 (AUC=0.821)
  - ROC Fold 4 (AUC=0.732)
  - Mean ROC (AUC=0.754, 95%CI [0.695-0.812])
  - Chance

- **b. Pathological image model**
  - ROC Fold 0 (AUC=0.714)
  - ROC Fold 1 (AUC=0.821)
  - ROC Fold 2 (AUC=0.839)
  - ROC Fold 3 (AUC=0.786)
  - ROC Fold 4 (AUC=0.768)
  - Mean ROC (AUC=0.786, 95%CI [0.747-0.824])
  - Chance

- **c. Combined model**
  - ROC Fold 0 (AUC=0.857)
  - ROC Fold 1 (AUC=0.732)
  - ROC Fold 2 (AUC=0.839)
  - ROC Fold 3 (AUC=0.714)
  - ROC Fold 4 (AUC=0.857)
  - Mean ROC (AUC=0.818, 95%CI [0.745-0.855])
  - Chance

**Summary**

- **a** Patient information model.
- **b** Pathological image model.
- **c** Combined model.

ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; std. dev., standard deviation.
Supplementary Figure 5 Kaplan–Meier curves for the progression-free survival of the high and low efficacy groups predicted using our predictive models without the censored patients for all folds. a Patient information model. b Pathological image model. c Combined model.