COLLEGE of AMERICAN PATHOLOGISTS

ARCHIVES
of Pathology & Laboratory Medicine

The DOI for this manuscript is doi: 10.5858/arpa.2022-0051-OA

Overcoming the Interobserver Variability in Lung Adenocarcinoma Subtyping

A Clustering Approach to Establish a Ground Truth for Downstream Applications

Kris Lami, MD; Andrey Bychkov, MD, PhD, FRCPath; Keitaro Matsumoto, MD, PhD; Richard Attanoos, MBBS, FRCPath; Sabina Berezowska, MD, PhD; Luka Brcic, MD, PhD; Alberto Cavazza, MD; John C. English, FRCPC; Alexandre Todorovic Fabro, MD, PhD; Kaori Ishida, MD; Yukio Kashima, MD; Brandon T. Larsen, MD, PhD; Alberto M. Marchevsky, MD; Takuro Miyazaki, MD, PhD; Shimpei Morimoto, PhD; Anja C. Roden, MD; Frank Schneider, MD; Mano Soshi, MD; Maxwell L. Smith, MD; Kazuhiro Tabata, MD, PhD; Angela M. Takano, MD; Kei Tanaka, MMedSci; Tomonori Tanaka, MD; Tomoshi Tsuchiya, MD, PhD; Takeshi Nagayasu, MD, PhD; Junya Fukuoka, MD, PhD

• ***Context.***—**The accurate identification of different lung adenocarcinoma histologic subtypes is important for determining prognosis but can be challenging because of overlaps in the diagnostic features, leading to considerable interobserver variability.**

Objective.—**To provide an overview of the diagnostic agreement for lung adenocarcinoma subtypes among pathologists and to create a ground truth using the clustering approach for downstream computational applications.**

Design.—**Three sets of lung adenocarcinoma histologic images with different evaluation levels (small patches, areas with relatively uniform histology, and whole slide images) were reviewed by 18 international expert lung pathologists. Each image was classified into one or several lung adenocarcinoma subtypes.**

Results.—**Among the 4702 patches of the first set, 1742 (37%) had an overall consensus among all pathologists. The overall Fleiss κ score for the agreement of all subtypes was 0.58. Using cluster analysis, pathologists were hierarchically grouped into 2 clusters, with κ scores of 0.588 and 0.563 in clusters 1 and 2, respectively. Similar results were obtained for the second and third sets, with fair-to-moderate agreements. Patches from the first 2 sets that obtained the consensus of the 18 pathologists were retrieved to form consensus patches and were regarded as the ground truth of lung adenocarcinoma subtypes.**

Conclusions.—**Our observations highlight discrepancies among experts when assessing lung adenocarcinoma subtypes. However, a subsequent number of consensus patches could be retrieved from each cluster, which can be used as ground truth for the downstream computational pathology applications, with minimal influence from interobserver variability.**

(*Arch Pathol Lab Med.* doi: 10.5858/arpa.2022-0051-OA)

Lung cancer is the leading cause of cancer-related deaths globally, accounting for 18% of all cancer deaths, and is the second most commonly diagnosed cancer in both sexes.[1] One of its histologic and most common subtypes is adenocarcinoma (ADC), representing more than 40% of

Accepted for publication June 22, 2022.

Supplemental digital content is available for this article. See text for hyperlink.

From the Departments of Pathology (Lami, K. Tanaka, Fukuoka) and Surgical Oncology (Matsumoto, Miyazaki, Tsuchiya, Nagayasu) and the Innovation Platform & Office for Precision Medicine (Morimoto), Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; the Department of Pathology, Kameda Medical Center, Kamogawa, Japan (Bychkov); the Department of Cellular Pathology, Cardiff University, Cardiff, United Kingdom (Attanoos); the Institute of Pathology, Centre Hospitalier Universitaire Vaudois, University of Lausanne, Lausanne, Switzerland (Berezowska); the Diagnostic and Research Institute of Pathology, Medical University of Graz, Graz, Austria (Brcic); the Unit of Pathologic Anatomy, Azienda USL/IRCCS di Reggio Emilia, Reggio Emilia, Italy (Cavazza); the Department of Pathology, Vancouver General Hospital, Vancouver, British Columbia, Canada (English); the Department of Pathology and Legal Medicine, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil (Fabro); the Department of Pathology, Kansai Medical University, Osaka, Japan (Ishida); the Department of Pathology, Hyogo Prefectural Awaji Medical Center, Sumoto, Japan (Kashima); the Department of Laboratory Medicine and Pathology, Mayo Clinic, Scottsdale, Arizona (Larsen, Smith); the Department of Pathology, Cedars-Sinai Medical Center, Los Angeles, California (Marchevsky); the Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota (Roden); the Department of Pathology and Laboratory Medicine, Emory University, Atlanta, Georgia (Schneider); BonBon Co, Ltd, Kyoto, Japan (Soshi); the Department of Pathology, Kagoshima University, Kagoshima, Japan (Tabata); the Department of Anatomical Pathology, Singapore General Hospital, Singapore, Singapore (Takano); and the Department of Diagnostic Pathology, Kobe University Hospital, Kobe, Japan (T. Tanaka).

Soshi is the CEO of BonBon Co, Ltd. The other authors have no relevant financial interest in the products or companies described in this article.

This paper is based on results obtained from a project, 19100971-0, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Corresponding author: Junya Fukuoka, MD, PhD, Department of Pathology, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan 1-7-1, Sakamoto, Nagasaki 852-8523, Japan (email: fukuokaj@nagasaki-u.ac.jp).

total lung cancer cases.[2] In 2015, the World Health Organization (WHO) adopted 5 main histologic patterns of lung ADC, which correspond to its subtypes, namely lepidic, acinar, papillary, micropapillary, and solid, and several variants comprising invasive mucinous ADC, colloid ADC, fetal ADC, and enteric-type ADC.[2] These subtypes were reproduced in the 2021 WHO classification of thoracic tumors.[3]

It is important to accurately recognize these architectural patterns of lung ADC because they have shown to be an independent prognostic factor for relapse-free and overall survival. Lepidic, acinar, and papillary subtypes have been found to have a better prognosis than micropapillary, solid, and invasive mucinous patterns.[4–6] However, numerous studies have shown that the classification of lung ADC patterns suffers from interobserver variability.[7–10] As suggested in one study, a novel method for distinguishing the histologic subtypes of ADC may be needed for a more accurate and reliable diagnosis.[8]

Cluster analysis is a method of identifying relevant subgroups of items through statistical analysis to divide a group of entities into more uniform and mutually exclusive groups based on their correlations.[11] Several studies using a cluster analysis approach have been conducted in the field of lung cancer, using genes or immunohistochemical expression of biomarkers as variables.[12–14] Lung ADC subtyping, with its known interobserver variability, can benefit from cluster analysis by refining the diagnostic criteria and obtaining the ground truth.

In recent years, digitization has emerged in the pathology domain, with whole slide image (WSI) evaluation of scanned glass slides entering the daily workflow of pathologists.[15,16] With this advancement, many convolutional neural networks (CNNs) have been developed to recognize and classify different tumor subtypes and tumor grading[17–19] and predict recurrence and gene mutation in lung cancer.[20,21] From the histologic point of view, several deep learning algorithms conceived to recognize certain subtypes of lung ADC have already been developed, with the training set based on annotated ground truths obtained from 1 to 3 pathologists.[22,23] Regarding these algorithms, little is known about the interobserver variability in the recognition of lung ADC subtypes. Creating a set of consensus images of lung ADC subtypes would decrease the interobserver variability. It can be further used as the ground truth to train CNNs for automatic detection and classification of lung ADC subtypes, improving their diagnostic accuracy.

In this regard, this study aimed to assess the agreement among expert pulmonary pathologists to diagnose lung ADC subtypes and obtain a reliable subset of images in agreement using a cluster approach.

MATERIALS AND METHODS

Study Design

This retrospective study used a series of cases from a single institute between 2007 and 2020. This study was approved by the Clinical Research Review Committee of the Nagasaki University Hospital (Nagasaki, Japan) (approval no. 20042008-2).

The overall workflow of the study is shown in Figure 1. Using the institutional electronic medical records, we retrospectively retrieved 191 representative surgically resected lung ADC cases encompassing all histologic subtypes from the Nagasaki University Hospital, Nagasaki, Japan. The selection criteria included a solitary ADC of any size. Cases with metastatic lesions and double lung cancer of any histologic subtype were excluded. Each glass slide was scanned with a ×20 objective (0.5 μm/pixel resolution) using Aperio Scanscope CS2 digital slide scanner (Leica Biosystems, Buffalo Grove, Illinois), which produced 330 WSIs. The selected cases were divided into 3 sets: the first set included 12 WSIs from 12 patients with lung ADC, including the 6 major histologic subtypes; the second set included 79 WSIs retrieved from 37 patients; and the third set included the remaining 239 WSIs from 142 patients.

Each set had a different evaluation method. For the first set, the dominant pattern of small patches was recorded. The 12 WSIs were segmented into 1-mm^2 patches with a 2 μm per pixel resolution, corresponding to a ×5 magnification, resulting in 4702 patches after excluding those with more than 80% blank background.

For the second set, the dominant subtype of annotated areas representing relatively uniform lung ADC histology was recorded. The 79 WSIs were annotated by 1 pathologist in training under the supervision of an expert pulmonary pathologist. The annotation was performed with the SplineAnnotation function of the Automated Slide Analysis Platform software, version 1.9 (ASAP, Computation Pathology Group, Nijmegen, The Netherlands).

For the third set, every subtype present in the entire slide of each of the 239 WSIs was recorded.

Seventeen expert pulmonary pathologists (Japan, 5; United States, 5; Austria, 1; Brazil, 1; Canada, 1; Italy, 1; Singapore, 1; Switzerland, 1; United Kingdom, 1) from different institutions with more than 15 years of experience on average and 1 pathologist in training evaluated the histologic subtype of the 4702 patches of the first set and the annotated areas of the 79 WSIs of the second set by assessing 1 lung ADC subtype to each element, based on the WHO classification.[2,3] The subtypes included lepidic, acinar, papillary, micropapillary, solid, invasive mucinous ADC, other cancer types (for morphology with a complex glandular pattern, cribriform pattern, colloid ADC, fetal ADC, or enteric ADC), and no carcinoma cells (for patches of the first set not visibly containing any cancer cells). Invasion was defined as a lung ADC subtype other than the lepidic growth.[2]

After the evaluation, annotated areas of the 79 WSIs of the second set were segmented into 1-mm^2 patches, producing 8554 patches. Patches of the first set were reviewed using an application for smartphones/tablets created explicitly for this purpose. The 79 WSIs of the second set were uploaded to the PathPresenter platform (PathPresenter.net, New York, New York), and a spreadsheet was created to help pathologists sort out each annotated area with the estimated lung ADC subtypes. For the third set, 15 pathologists from the same group (3 pathologists were not available to join this step of the study) were asked to provide a case-level diagnosis of the 142 patients (representing 239 WSIs) by determining the dominant and minor subtypes of lung ADC for each patient and their estimated percentages in 5% increments.

Staining

Because of the fading of stains, old glass slides were washed out with 1% acid alcohol and then restained with hematoxylin-eosin using the Tissue-Tek Prisma Plus Automated Slide Stainer (Sakura Finetek, Tokyo, Japan).

Statistical Analysis

The Cohen κ coefficient was used to evaluate the pairwise agreement for invasive cancer versus noninvasive cancer patches and cancer versus noncancer patches, with a total of 153 κ values calculated for each category from different combinations of the 18 pathologists. To evaluate the overall and histologic subtype agreements from multiple raters, the Fleiss κ coefficient was calculated. Agreements were defined as poor, slight, fair, moderate, substantial, and almost perfect for κ values of less than 0, 0.01 to 0.20, 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80, and 0.81 to 1.0, respectively.

Using the Ward method, hierarchical cluster analysis was performed with the R software (R Foundation for Statistical Computing, Vienna, Austria) from labels of the 4702 patches of the 18 pathologists, and these labels were considered as categorical

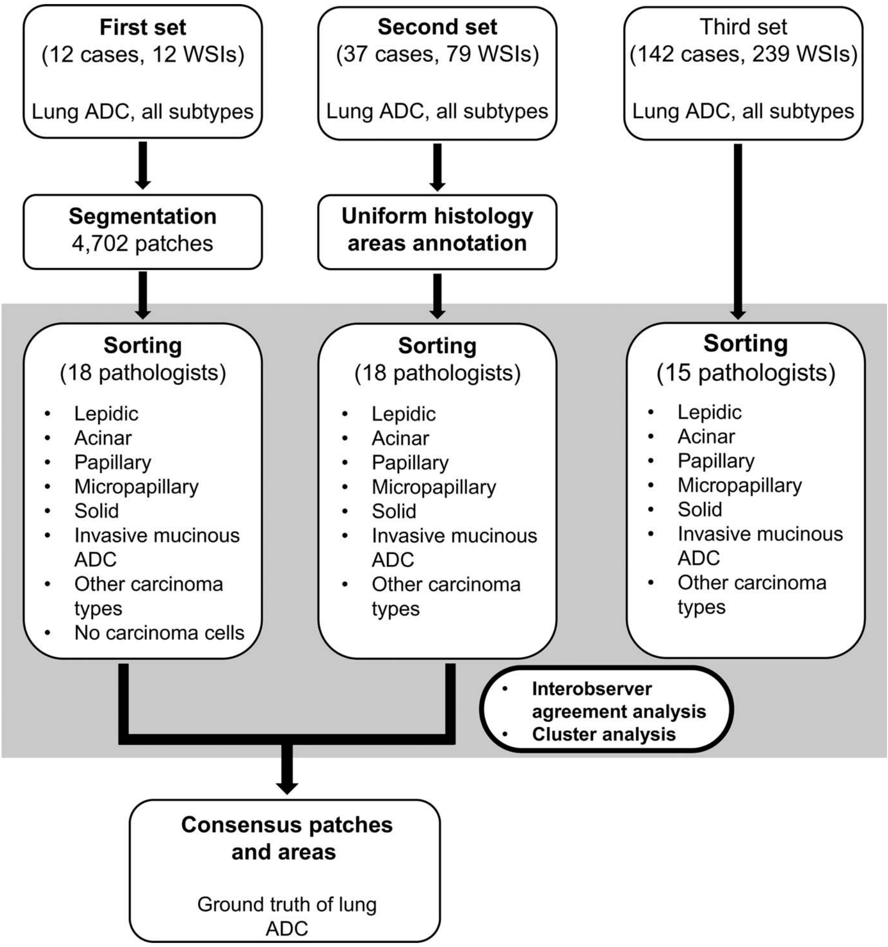

Figure 1. *Flowchart of the study. Abbreviations: ADC, adenocarcinoma; WSI, whole slide image.*

variables. The distance between pathologists was determined by the Cramér V, calculated using the vcd package.[24] The uncertainty in the result from the clustering analysis was assessed via multiscale bootstrap resampling.[25]

RESULTS

Agreements of the First Set

Lung ADC histologic subtypes were assigned to the first set of 4702 patches by the 18 pathologists. The overall proportion of the selected histologic patterns by each pathologist is shown in Figure 2, A, with scored patches for each pathologist shown in Supplemental Figure 1 (see supplemental digital content). The pathologists selected one of the lung ADC histologic patterns for an average percentage of 52.3%, and the average percentage of the total patches considered as containing no cancer cells was 47.7%.

Among the 4702 patches analyzed at ×5 magnification, all 18 pathologists had a complete consensus on 1742 patches (37%), including 1520 patches labeled as no carcinoma cells (Figure 2, B). After the no carcinoma cells patches were excluded, the pathologists had a consensus on 222 patches (4.7% of total patches) labeled as one of the lung ADC subtypes. Solid pattern was the subtype with the most consensus patches, with all 18 pathologists having complete agreement on 180 patches, followed by invasive mucinous (29 patches), acinar (8 patches), and micropapillary (5 patches) subtypes. The 18 pathologists had no complete consensus on lepidic, papillary, and other carcinoma types (Figure 2, C).

Pairwise agreements for invasive versus noninvasive patches were evaluated with Cohen κ score, varying from 0.05 to 0.76 (Supplemental Table 1). Agreements for cancer versus noncancer patches were also evaluated, with a narrower interval than the invasive versus noninvasive patches. The highest score of 0.93 was observed between pathologists 15 and 16 (Supplemental Table 2).

Interobserver variability for the 12 cases of the first set was highlighted with individual heat maps, when patches with their respective lung ADC patterns given by each pathologist were superimposed on the original WSIs. In 1 case, there was a disagreement among pathologists about whether the case was predominantly invasive or noninvasive. In another case, the heat map showed the WSI being recognized by different pathologists as acinar predominant, papillary predominant, or even invasive mucinous ADC predominant (Figure 3).

Pathologists' Grouping in Clusters

Based on the assessment of the uncertainty in the results obtained from the first set, the pathologists were grouped into 2 clusters, and 3 pathologists were outlying from both clusters (Figure 4). Clusters 1 and 2 consisted of 10 and 5 pathologists, respectively. The main difference between the 2 clusters was predominantly seen in 1 case, where pathologists from cluster 1 agreed for an invasive mucinous–predominant case, whereas pathologists from cluster 2 opted for a micropapillary or other cancer types–predominant case (Supplemental Figure 2, A). The overall

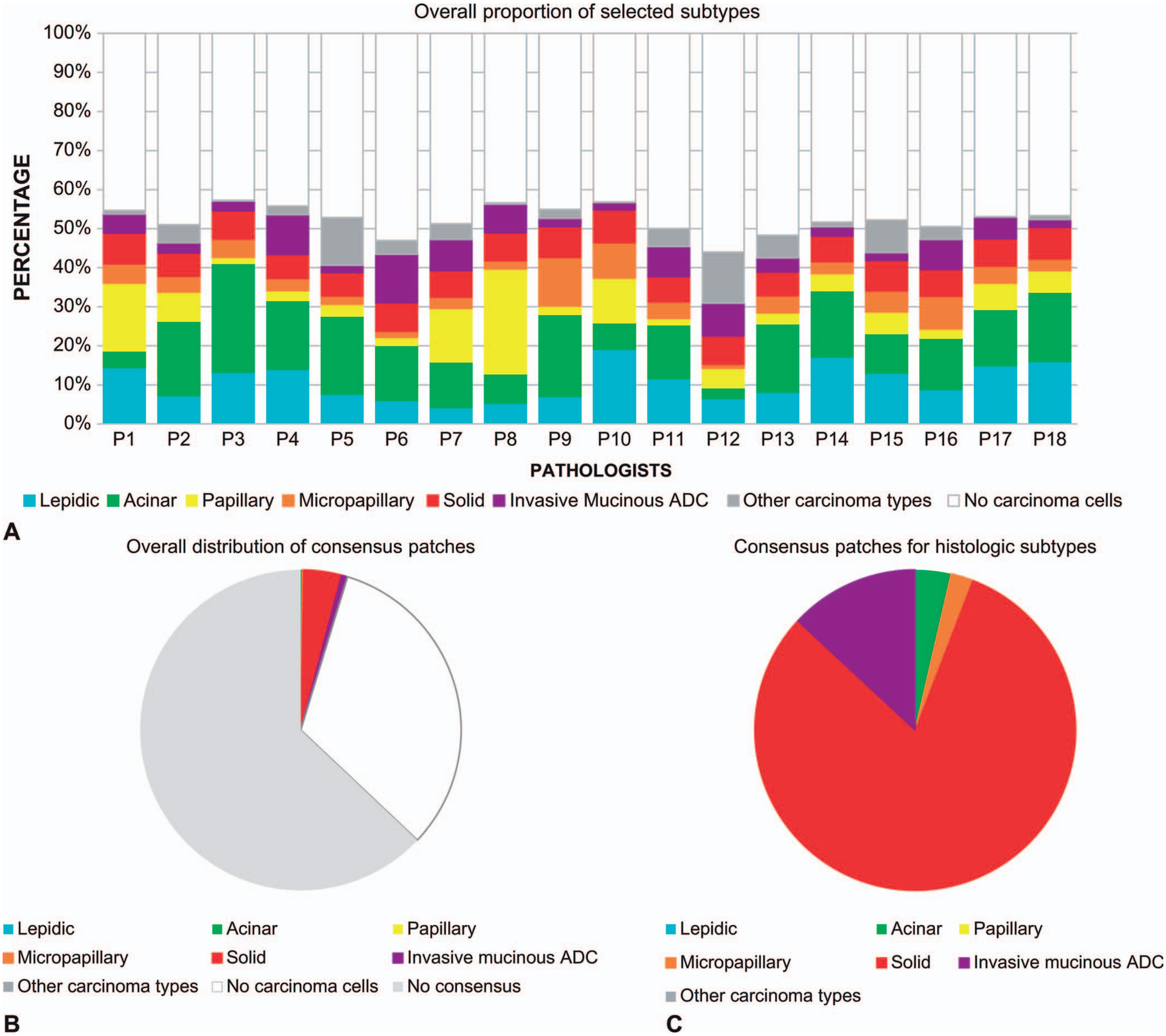

Figure 2. *Overview of the first set. A, Overall proportion of the selected patterns by the 18 pathologists. B, Distribution of the consensus patches for the 18 pathologists, wherein 37% of patches had a complete agreement, including 1520 patches labeled as "no carcinoma cells." C, Proportion of the consensus patches for each histologic subtype (with "no carcinoma cells" removed). The solid subtype had the most consensus, accounting for 81% of all histologic subtypes. Abbreviations: ADC, adenocarcinoma; P, pathologist.*

agreement (all lung ADC subtypes and "no carcinoma cells" patches combined) of the 18 pathologists calculated with the Fleiss κ score was 0.585. Pathologists from clusters 1 and 2 had overall agreements of 0.588 and 0.563 within each cluster, respectively. The highest agreements among the 18 pathologists and for both clusters were seen for categories separating cancer and noncancer patches as well as solid and other subtypes patches, both of them achieving an almost perfect agreement, followed by invasive mucinous ADC and other subtypes, with a substantial to an almost perfect agreement (Table 1).

Agreements of the Second and Third Sets

All 18 pathologists attributed lung ADC histologic patterns to the annotated area of the 79 WSIs from the 37 cases of the second set. The pathologists' agreements using the Fleiss κ score were calculated again. The κ values showed scores similar to but slightly lower than those obtained from the first set (Table 2). Once again, the solid subtype achieved the highest score but with only a substantial agreement among the pathologists ($\kappa = 0.792$). The lowest agreement was observed for the other carcinoma type pattern, with a κ score of 0.229, indicating a slight agreement.

Fifteen pathologists were then asked to label the 142 cases of the third set, comprising 239 WSIs with a lung ADC subtype. Labeling was done on WSIs on a case-level basis by determining the dominant and minor subtypes and estimating the percentage of every subtype present in the case, in 5% increments. The overall agreement of the 15

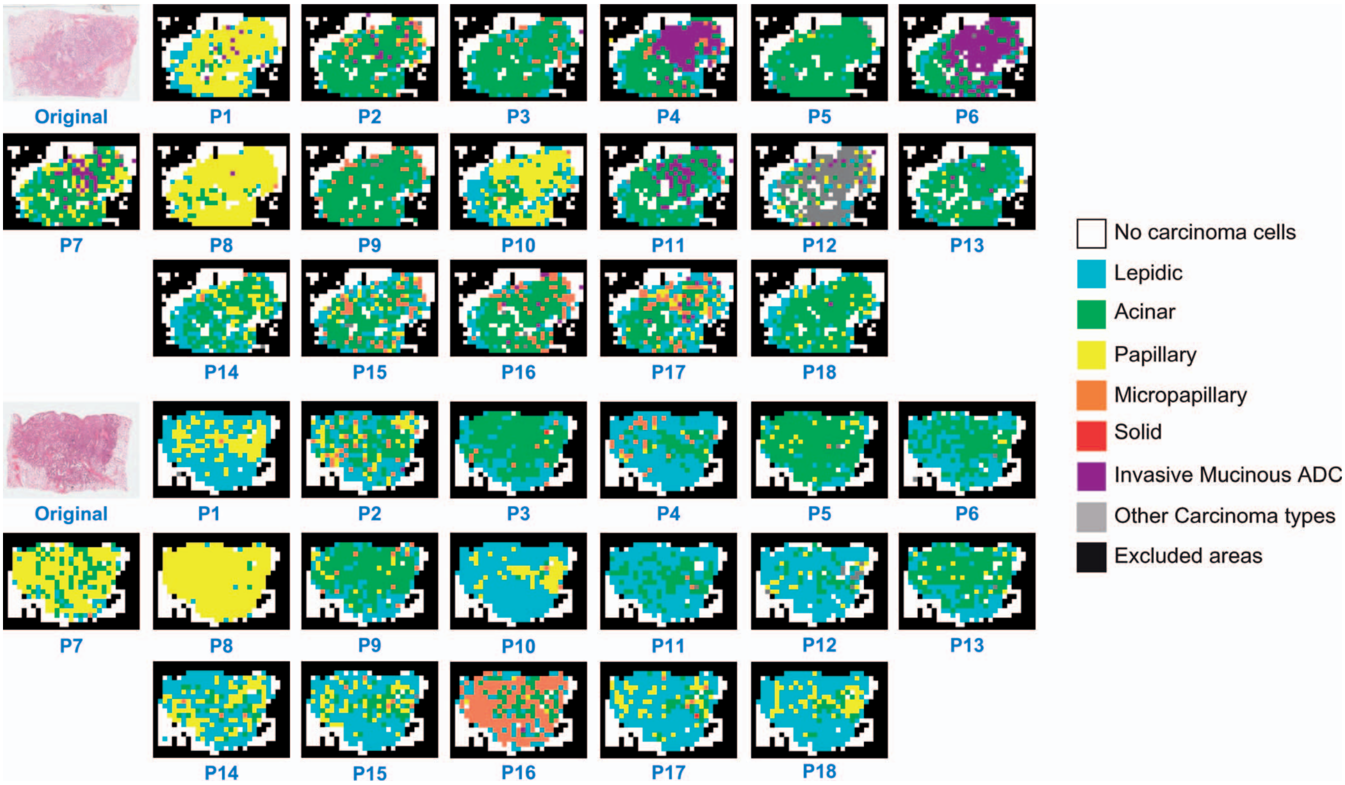

Figure 3. *Individual heat maps of whole slide images, with each color representing a specific pattern. The top image is predominantly acinar, papillary, or invasive mucinous adenocarcinoma debatable among pathologists. On the bottom, the case being predominantly noninvasive or invasive adenocarcinoma was debatable. Abbreviations: ADC, adenocarcinoma, P, pathologist.*

pathologists for the predominant subtype achieved a κ score of 0.338, indicating a fair agreement, which is the lowest of the 3 sets. The highest agreement was observed for invasive mucinous ADC, with a substantial agreement (κ = 0.763) (Table 2).

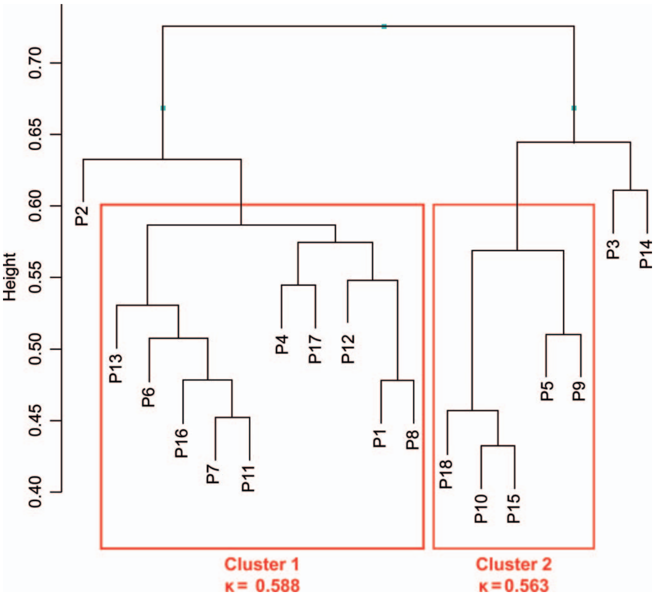

Figure 4. *Hierarchical clustering of the 18 pathologists. Two clusters were created, with the overall Fleiss κ score of each cluster indicated. Abbreviation: P, pathologist.*

Evaluation of Invasive Morphology by the 2 Clusters

Next, the case-level diagnoses of the third set given by the 15 pathologists were used to evaluate the lung ADC tumor grades, as proposed by the International Association for the Study of Lung Cancer pathology committee[26] and adopted by the WHO.[3] We evaluated 133 cases from the third set after excluding cases assessed as invasive mucinous ADC by 1 of the 15 pathologists. After excluding outlier pathologists, we retained 12 pathologists for the evaluation. Survival analysis was conducted to evaluate the ability of pathologists and clusters to separate noninvasive-predominant from invasive-predominant tumors. Noninvasive-predominant tumors were defined as grade 1 tumors according to the aforementioned lung ADC grading system, and invasive-predominant tumors were defined as grade 2 or grade 3 tumors. The patients' demographic characteristics are listed in Supplemental Table 3. The plotted Kaplan-Meier curves showed statistical significance for both clusters, with *P* values of .03 and .02 for clusters 1 and 2, respectively (Figure 5, A and B). Cluster 1 showed a better significance than 4 pathologists of its cluster (pathologist [P] 6, P7, P4, and P8), whereas cluster 2 showed a better stratification than 3 pathologists of its cluster (P18, P10, and P9), outperformed by only 1 pathologist (P5). These results show that the clustering approach provided a reasonable assessment in the prediction of invasive morphology.

Consensus Patches

The first and second sets spawned several patches or areas with high agreements among pathologists. By defining consensus patches as those attributed 6 of 10 identical labels

Table 1. Fleiss κ Scores for Different Agreements of the Lung Adenocarcinoma Subtypes[a]

Agreement Category	All 18 Pathologists	Cluster 1	Cluster 2
Overall	0.585	0.588	0.563
Cancer versus no cancer	0.844	0.836	0.811
Invasive versus noninvasive	0.661	0.610	0.608
Acinar versus other patterns	0.572	0.448	0.499
Papillary versus other patterns	0.610	0.364	0.377
Acinar + papillary versus other patterns	0.562	0.475	0.442
Micropapillary versus other patterns	0.713	0.580	0.689
Solid versus other patterns	0.902	0.900	0.917
Invasive mucinous versus other patterns	0.794	0.904	0.808
Other carcinomas versus other patterns	0.673	0.500	0.541

[a] <0, no agreement; 0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.0, almost perfect agreement.

Table 2. Fleiss κ Scores of the First, Second, and Third Sets for Agreements of Various Categories[a]

	Set		
Agreement Category	First	Second	Third
Overall	0.585	0.522	0.338
Cancer versus no cancer	0.844	. . .	. . .
Invasive versus noninvasive	0.661	0.608	0.453
Acinar versus other patterns	0.572	0.392	0.271
Papillary versus other patterns	0.610	0.376	0.158
Acinar + papillary versus other patterns	0.562	0.484	0.369
Micropapillary versus other patterns	0.713	0.679	0.320
Solid versus other patterns	0.902	0.792	0.635
Invasive mucinous versus other patterns	0.794	0.597	0.763
Other carcinomas versus other patterns	0.673	0.229	0.053

[a] <0, no agreement; 0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.0, almost perfect agreement.

for cluster 1 and 3 of 5 identical labels for cluster 2, we were able to retrieve a subsequent number of consensus patches from the 2 clusters (Table 3). For the first set, a total of 3733 and 4214 consensus patches were retrieved from clusters 1 and 2, respectively. The 2 clusters had 3529 common patches, including 3312 sharing the same lung ADC subtype label (Figure 6, A). In total, 217 patches had different subtype labels between the 2 clusters (Supplemental Figure 2, B). For the second set, 6409 patches of 61 WSIs and 7188 patches of 66 WSIs met consensus for clusters 1 and 2, respectively, after the segmentation of annotated areas. The 2 clusters had 55 common WSIs (5853 patches), including 50 sharing the same lung ADC histologic subtype label (5285 patches) (Figure 6, B). Only 5 WSIs (568 patches) had different subtype labels between the 2 clusters (Supplemental Figures 2, C, and 3, A through E). We considered the resulting consensus patches as the ground truth of lung ADC subtypes (Table 3).

DISCUSSION

This study investigated the interobserver agreement of lung ADC subtypes among expert pulmonary pathologists from different institutions. To date, this is the first study to compare an interobserver agreement using different evaluation levels and cluster analysis for determining lung ADC subtypes. Based on their evaluation, 2 distinct clusters of pathologists were created, with a subsequent number of consensus patches resulting from both clusters. These patches can help set a ground truth for lung ADC subtypes, which can be further used to train deep learning algorithms.

Overall, the κ score for the agreement of histologic patterns in the first set was 0.585, with a specific agreement for each lung ADC subtype, ranging from 0.562 to 0.902, indicating a moderate to an almost perfect agreement (only achieved for the solid subtype). These results are similar to previous studies focused on the interobserver variability of lung ADC subtypes, with the κ score ranging from fair to substantial.[7,8,10,27] Shih et al[8] found a significant difference in assessing the invasive size of the tumor, with up to a 19-mm difference among pathologists in one case. This highlights the relative difficulty of correctly setting tumor invasion, as seen in our study with the substantial agreement in differentiating invasive and noninvasive patches, with a Fleiss κ score of 0.66. The pairwise interobserver agreement for invasive versus noninvasive patterns had a wide-ranging Cohen κ score, from 0.05 to 0.763. This can be problematic in clinical practice because recognizing the noninvasive/lepidic component of cancer and the size of the invasive component are determinants of the T stage of lung ADC, with a direct impact on the disease-free survival in cases of ADC in situ, minimally invasive ADC, and lepidic-predominant ADC, and on the prognosis when determining the invasive size.[28] This scenario is best illustrated with personal heat maps. In some cases, it was debatable whether the diagnosis was primarily lepidic predominant or invasive pattern predominant (Figure 3).

The interobserver variability for cancer versus noncancer patches had a Cohen κ score ranging from 0.65 to 0.92, indicating substantial to almost perfect agreement. For the overall agreement, the κ score was 0.84, indicating that there was a consensus when differentiating the patches containing tumor cells from the ones that showed benign structures. One would assume a perfect agreement for differentiating cancer from noncancerous patches. The relatively low agreement was probably due to certain difficulties in distinguishing in some patches cancer cells from other floating cells, such as macrophages or reactive epithelial cells (Supplemental Figure 4, A through O). The low agreement can also be explained by pathologists' fatigue, as they had to sort a considerable number of patches. Another explanation is that some patches contained very few tumor cells compared with the predominant nontumor area, which some pathologists described as no carcinoma cells, whereas others indicated the presence of tumor cells by attributing the corresponding subtype.

In this study, the solid pattern had the most consensus patches, with the highest agreement among lung ADC subtypes, similar to that reported by other authors.[7,27] The agreement of solid pattern was also high in the frozen specimen section analyzed in one study,[29] showing the relative ease of distinguishing this subtype from others.

Acinar and papillary patterns had the lowest agreements among lung ADC subtypes in the first set. Overall, these 2 patterns had moderate to substantial agreements, with Fleiss κ scores of 0.57 and 0.61, respectively. A possible reason for the low agreement can be the relative resem-

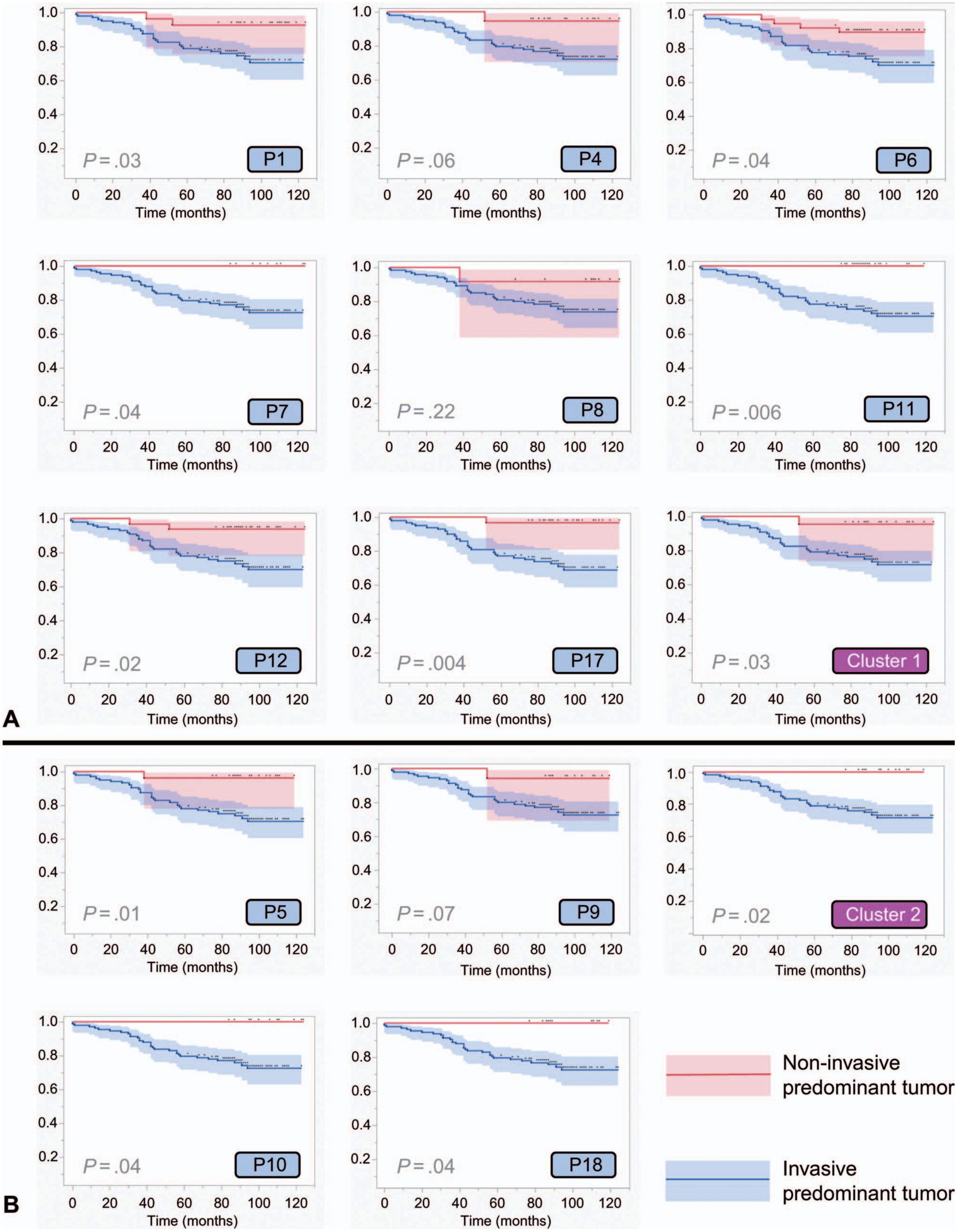

Figure 5. *Overall survival of 133 cases of the third set stratified by invasive-predominant and noninvasive-predominant tumors. Both clusters achieved significance, with superior performance compared with some pathologists. A, Kaplan-Meier curves for pathologists from cluster 1 and cluster 1 consensus. B, Kaplan-Meier curves for pathologists from cluster 2 and cluster 2 consensus. Abbreviation: P, pathologist.*

Table 3. Number of Consensus Patches for Each Set and Cluster

	Cluster 1								Cluster 2							
	L	A	P	MP	S	IM	O	N	L	A	P	MP	S	IM	O	N
First set	308	354	116	59	319	321	34	2222	551	596	154	192	371	90	84	2176
Second set	552	1507	1475	549	866	1460	0	NA	619	1803	2066	549	1034	1117	0	NA
Total	**860**	**1861**	**1591**	**608**	**1185**	**1781**	**34**	**2222**	**1170**	**2399**	**2220**	**741**	**1405**	**1207**	**84**	**2176**

Abbreviations: A, acinar; IM, invasive mucinous adenocarcinoma; L, lepidic; MP, micropapillary; N, no carcinoma cells; NA, not applicable; O, other carcinoma types; P, papillary; S, solid.

blance of those 2 patterns with other subtypes in equivocal cases. By definition, both acinar and papillary subtypes are characterized by glandular growth as the major component, with the difference being the presence of central fibrovascular cores in the papillary subtype. Lepidic growth with the alveolar spaces filled with papillary structures is also called papillary subtype, and micropapillary growth is characterized by papillary tufts without fibrovascular cores.[30] In some cases, it can be difficult to correctly identify fibrovascular cores; thus, papillary growth can be mistaken for other subtypes, and vice versa. Moreover, the interobserver agreement did not improve when merging acinar and

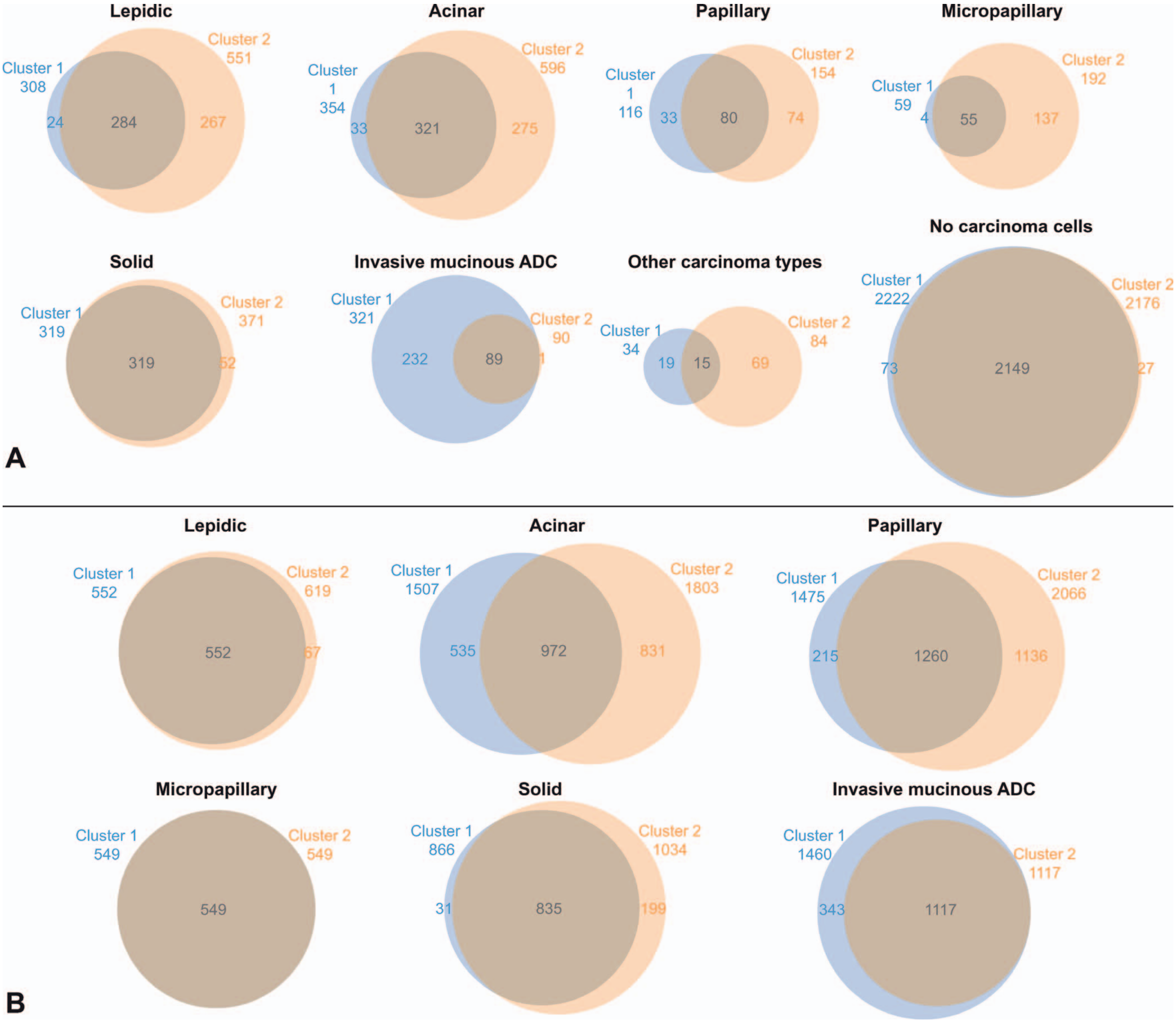

Figure 6. *Consensus patches from clusters 1 and 2 and common patches from both clusters sharing the same lung ADC labels. Numbers outside the circles represent the total number of consensus patches from one cluster. A, Venn diagrams showing the number of consensus patches (intersection of the 2 circles) from the first set sharing the same label in both clusters. B, Venn diagrams showing the number of consensus patches (intersection of the 2 circles) from the second set sharing the same label in both clusters. Abbreviation: ADC, adenocarcinoma.*

papillary patterns and comparing them with other patterns (Table 1), indicating difficulty in distinguishing these 2 patterns from other patterns.

These low agreement values could arguably be due to the examination techniques. In the first set, pathologists evaluated lung ADC subtypes on small patches, limiting the appreciation of the surrounding area to assess a subtype correctly. We therefore provided pathologists a second set containing 79 WSIs for sorting out the histologic patterns of the annotated area, with relatively uniform histology. They had the freedom to magnify each slide up to ×20 magnification, with close scrutiny of histologic architecture. Again, the papillary and acinar patterns showed low agreement, with κ scores of 0.392 and 0.376, respectively, indicating fair agreement. In the second set, the "other carcinoma" subtype showed a worse agreement than that obtained from the first set, with a Fleiss κ score of 0.229, indicating a slight agreement. A third set with a full whole slide evaluation was also provided, with the determination of dominant and minor subtypes and their percentages. Similar trends for the κ score were again seen, with the acinar and papillary subtypes achieving slight and fair agreements, respectively. Agreements of the third set were slightly lower than the second set. The "other carcinoma" subtype showed the worst agreement among the lung ADC subtypes in the third set. This is probably due to the misrecognition of cribriform or complex glandular patterns as "other carcinoma," as recommended in this study. These patterns were regarded as acinar ADC and have been recently described as high-grade patterns.[3,31] Some pathologists may have failed to correctly identify these patterns, thus the slight agreement seen on the third set for the "other carcinoma" subtype (Supplemental Figure 5, A through H). Another reason that may explain the low agreement is the possible intraobserver variability. Supplemental Figure 2, B and C, showing the overlapping consensus patches displaying different labels between the 2 clusters, revealed a discrepancy in labels of discordant consensus patches between the first and second sets. This may be explained by a probable change in the diagnostic criteria between the sets for a given pathologist, further decreasing the agreement of different sets.

Previous studies[7,9,10,27,32–35] focusing on interobserver variability of non–small cell lung cancer, lung ADC, or lung ADC subtypes used conventional glass slides and microphotographs/still images to evaluate the pathologists' agreements. A single study[8] used WSIs to investigate the agreement for the classification of ADC in situ, minimally invasive ADC, and invasive ADC of the lung. We believe that our study is the first to assess the agreement for lung ADC subtypes using WSIs. Different evaluation methods showed that the agreements obtained with close scrutiny of small patches (first set) are superior to the ones obtained with analyses of an enclosed area in a WSI (second set) or the entire WSI (third set) (Table 2). These elements show that inspection of small areas with minimal morphologic features on a low power results in better agreements when evaluating lung ADC subtypes. It is, however, important to note that the difference in the evaluation method between the first set (inspection of small patches) and the 2 other sets (WSI examination) may have introduced some variable that affected pathologists' performance in the lung ADC subtyping beyond the change in the size of the field reviewed, thus the difference in agreement values between sets.

In this study, we introduced the cluster analysis technique in the evaluation of lung ADC subtypes. The selection of the clustering approach over the majority rule is explained by the fact that the cluster analysis creates subgroups of pathologists with distinctive diagnostic criteria for lung ADC subtypes, which can help to refine ground truth images. Also, an 80% agreement rule was abandoned as it resulted in few consensus patches. The pathologists' answers from the first set created 2 hierarchical clusters, with the main difference seen principally in 1 of the 12 cases included for evaluating the pathologists' agreement (Supplemental Figure 2, A). Cluster 1 agreed with an invasive mucinous ADC–predominant cancer, whereas cluster 2 favored micropapillary and other cancer subtypes. Although these patterns are considered high-grade tumors,[36–38] the presence of micropapillary features alone is a factor of poor prognosis for lung cancer.[39–41] Moreover, invasive mucinous ADC has a different genetic signature,[30] making the distinction more important.

The clustering approach has been evaluated by its ability to predict overall survival based on the predominance of invasive or noninvasive features of a given tumor. For this matter, the recently adopted lung ADC grading system has been applied to pathologists' answers to the third set, and consensus has been obtained for the 2 clusters. The resulting Kaplan-Meier curves showed a better stratification of invasive-predominant and noninvasive-predominant tumors, cluster 2 surpassing the majority of pathologists for the particular cohort used in this study. We therefore believe that consensus patches derived from these 2 clusters can help in accurate recognition of lung ADC subtypes and, by extension, assessment of invasion, which can be challenging and problematic in some cases.[42]

The first and second sets produced a consequent number of consensus patches from the 2 clusters, which shared up to 88% of identically labeled patches (Figure 6, A and B). We considered these patches as the ground truth of lung ADC subtypes. These patches will be used as training sets for our future study to develop deep learning algorithms. Patches will be augmented to increase the amount of the training set, with the third set serving as a testing set for the algorithms. Survival curves obtained from the deep learning algorithms can be compared with those obtained from the pathologists for evaluation. After the validation with survival analysis, the whole set of ground truth images will be made publicly accessible.

This study had some limitations. First, only 49 cases from a single institute were used to obtain consensus patches. Although every common invasive nonmucinous lung ADC subtype and invasive mucinous ADC cases were included, other ADCs, such as colloid ADC or fetal ADC, among others, were not sufficiently represented. Second, the annotation of the second set was done by a trainee pathologist. Although the process was supervised by an expert pulmonary pathologist, histologic homogeneity of the annotated area cannot be completely guaranteed, thus raising the possibility of increasing the interobserver variability. However, the agreement of the third set revealed a similar score to that obtained from the second set. This shows that the agreement was still low, even if pathologists had the freedom to evaluate the WSIs without limitations. Third, the consensus patches retrieved from the 2 clusters were not demonstrated to improve the overall agreement. Although we showed that the clustering approach resulted in better identification of invasion than some pathologists alone, consensus patches themselves were not used to evaluate the improvement of lung ADC subtype recognition. This issue needs to be addressed in a separate study, in which consensus patches can be used to train CNN models.

Nevertheless, our study proved the possibility of achieving a consensus on ascertaining lung ADC subtypes despite high interobserver variability, which is characteristic of lung ADC subtyping. Moreover, this is the first study to evaluate this agreement using cluster analysis, which can refine the diagnostic criteria of lung ADC.

CONCLUSIONS

The agreement of pathologists for the determination of lung ADC subtypes varied from slight to almost perfect using the process of examining patches at low power, annotating the area on a WSI, and performing case-level diagnosis using WSIs. Despite the existing interobserver variability, a subsequent number of consensus patches could be retrieved for each lung ADC subtype. These patches result from the general agreement of pulmonary pathologists and thus can be considered as the ground truth of lung ADC subtyping to be further used to train CNNs for automatic recognition of different lung ADC subtypes.

The authors would like to express their gratitude to Chikako Furukawa and Ai Mori for their help in the restaining of faded glass slides and immunostaining. The authors would also like to thank Tatsuro Ibe, MInf, of BonBon Co, Ltd, for his conception of the mobile application and heat maps.

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. Travis WD, Brambilla E, Burke AP, et al. *WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart*. 4th ed. Lyon, France: International Agency for Research on Cancer; 2015. *WHO Classification of Tumours*; vol 7.

3. WHO Classification of Tumours Editorial Board. *Thoracic Tumours*. 5th ed. Lyon, France: International Agency for Research on Cancer; 2021. *WHO Classification of Tumours*; vol 5.

4. Motono N, Matsui T, Machida Y, Usuda K, Uramoto H. Prognostic significance of histologic subtype in pStage I lung adenocarcinoma. *Med Oncol*. 2017;34(6):100.

5. Zhao X, Zhang Y, Qian K, Zhao L, Wang W, Teng LH. Prognostic significance of the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification of stage I lung adenocarcinoma: a retrospective study based on analysis of 110 Chinese patients. *Thorac Cancer*. 2017;8(6):565–571.

6. Eguchi T, Kadota K, Park BJ, Travis WD, Jones DR, Adusumilli PS. The new IASLC-ATS-ERS lung adenocarcinoma classification: what the surgeon should know. *Semin Thorac Cardiovasc Surg*. 2014;26(3):210–222.

7. Warth A, Stenzinger A, Von Brünneck AC, et al. Interobserver variability in the application of the novel IASLC/ATS/ERS classification for pulmonary adenocarcinomas. *Eur Respir J*. 2012;40(5):1221–1227.

8. Shih AR, Uruga H, Bozkurtlar E, et al. Problems in the reproducibility of classification of small lung adenocarcinoma: an international interobserver study. *Histopathology*. 2019;75(5):649–659.

9. Boland JM, Wampfler JA, Yang P, Yi ES. Growth pattern-based grading of pulmonary adenocarcinoma—analysis of 534 cases with comparison between observers and survival analysis. *Lung Cancer*. 2017;109:14–20.

10. Wright J, Churg A, Kitaichi M, Yang HM, Hyde D, Yi E. Reproducibility of visual estimation of lung adenocarcinoma subtype proportions. *Mod Pathol*. 2019;32(11):1587–1592.

11. Hair JF, Black WC, Babin BJ, Anderson RE. *Multivariate Data Analysis*. 8th ed. Andover, United Kingdom: Cengage Learning EMEA; 2019.

12. Fei H, Chen S, Xu C. Interactive verification analysis of multiple sequencing data for identifying potential biomarker of lung adenocarcinoma. *Biomed Res Int*. 2020;2020:8931419.

13. Hammer SH, Prall F. Close relation of large cell carcinoma to adenocarcinoma by hierarchical cluster analysis: implications for histologic typing of lung cancer on biopsies. *Appl Immunohistochem Mol Morphol*. 2015;23(8):550–557.

14. Sterlacci W, Fiegl M, Juskevicius D, Tzankov A. Cluster analysis according to immunohistochemistry is a robust tool for non-small cell lung cancer and reveals a distinct, immune signature-defined subgroup. *Appl Immunohistochem Mol Morphol*. 2020;28(4):274–283.

15. Hanna MG, Parwani A, Sirintrapun SJ. Whole slide imaging: technology and applications. *Adv Anat Pathol*. 2020;27(4):251–259.

16. Sakamoto T, Furukawa T, Lami K, et al. A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. *Transl Lung Cancer Res*. 2020;9(5):2255–2276.

17. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21(2):233–241.

18. Zhang H, Mo J, Jiang H, et al. Deep learning model for the automated detection and histopathological prediction of meningioma. *Neuroinformatics*. 2021;19(3):393–402.

19. Couture HD, Williams LA, Geradts J, et al. Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer*. 2018;4(1):30.

20. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–1567.

21. Wang X, Janowczyk A, Zhou Y, et al. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci Rep*. 2017;7(1):13543.

22. Gertych A, Swiderska-Chadaj Z, Ma Z, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci Rep*. 2019;9(1):1483.

23. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep*. 2019;9(1):3358.

24. Meyer D, Zeileis A, Hornik K. vcd: visualizing categorical data [computer program]. R package version 1.4-9. https://cran.r-project.org/web/packages/vcd/vcd.pdf2021. Accessed January 1, 2022.

25. Suzuki R, Terada Y, Shimodaira H. pvclust: hierarchical clustering with p-values via multiscale bootstrap resampling [computer program]. R package version 2.2-0. https://github.com/shimo-lab/pvclust. Published 2019. Accessed January 1, 2022.

26. Moreira AL, Ocampo PSS, Xia Y, et al. A grading system for invasive pulmonary adenocarcinoma: a proposal from the International Association for the Study of Lung Cancer Pathology Committee. *J Thorac Oncol*. 2020;15(10):1599–1610.

27. Thunnissen E, Beasley MB, Borczuk AC, et al. Reproducibility of histopathological subtypes and invasion in pulmonary adenocarcinoma: an international interobserver study. *Mod Pathol*. 2012;25(12):1574–1583.

28. Travis WD, Asamura H, Bankier AA, et al. The IASLC lung cancer staging project: proposals for coding T categories for subsolid nodules and assessment of tumor size in part-solid tumors in the forthcoming eighth edition of the TNM classification of lung cancer. *J Thorac Oncol*. 2016;11(8):1204–1223.

29. Yeh Y-C, Nitadori J, Kadota K, et al. Using frozen section to identify histologic patterns in stage I lung adenocarcinoma $\leq$ 3 cm: accuracy and interobserver agreement. *Histopathology*. 2015;66(7):922–938.

30. Travis WD, Brambilla E, Noguchi M, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol*. 2011;6(2):244–285.

31. Nicholson AG, Tsao MS, Beasley MB, et al. The 2021 WHO classification of lung tumors: impact of advances since 2015. *J Thorac Oncol*. 2022;17(3)362–387.

32. Warth A, Cortis J, Fink L, et al. Training increases concordance in classifying pulmonary adenocarcinomas according to the novel IASLC/ATS/ERS classification. *Virchows Arch*. 2012;461(2):185–193.

33. Wang C, Durra HY, Huang Y, Manucha V. Interobserver reproducibility study of the histological patterns of primary lung adenocarcinoma with emphasis on a more complex glandular pattern distinct from the typical acinar pattern. *Int J Surg Pathol*. 2014;22(2):149–155.

34. Boland JM, Froemming AT, Wampfler JA, et al. Adenocarcinoma in situ, minimally invasive adenocarcinoma, and invasive pulmonary adenocarcinoma—analysis of interobserver agreement, survival, radiographic characteristics, and gross pathology in 296 nodules. *Hum Pathol*. 2016;51:41–50.

35. Grilley-Olson JE, Hayes DN, Moore DT, et al. Validation of interobserver agreement in lung cancer assessment: hematoxylin-eosin diagnostic reproducibility for non-small cell lung cancer: the 2004 World Health Organization classification and therapeutically relevant subsets. *Arch Pathol Lab Med*. 2013;137(1):32–40.

36. Sica G, Yoshizawa A, Sima CS, et al. A grading system of lung adenocarcinomas based on histologic pattern is predictive of disease recurrence in stage I tumors. *Am J Surg Pathol*. 2010;34(8):1155–1162.

37. Yoshizawa A, Motoi N, Riely GJ, et al. Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. *Mod Pathol*. 2011;24(5):653–664.

38. Kadota K, Kushida Y, Kagawa S, et al. Cribriform subtype is an independent predictor of recurrence and survival after adjustment for the eighth edition of TNM staging system in patients with resected lung adenocarcinoma. *J Thorac Oncol*. 2019;14(2):245–254.

39. Wang W, Hu Z, Zhao J, et al. Both the presence of a micropapillary component and the micropapillary predominant subtype predict poor prognosis after lung adenocarcinoma resection: a meta-analysis. *J Cardiothorac Surg*. 2020; 15(1):154.

40. Cha MJ, Lee HY, Lee KS, et al. Micropapillary and solid subtypes of invasive lung adenocarcinoma: clinical predictors of histopathology and outcome. *J Thorac Cardiovasc Surg*. 2014;147(3):921–928.e2.

41. Lee G, Lee HY, Jeong JY, et al. Clinical impact of minimal micropapillary pattern in invasive lung adenocarcinoma: prognostic significance and survival outcomes. *Am J Surg Pathol*. 2015;39(5):660–666.

42. Borczuk AC. Updates in grading and invasion assessment in lung adenocarcinoma. *Mod Pathol*. 2022;35(suppl 1):28–35.