Histopathology

Histopathology 2019, **75**, 649–659. DOI: 10.1111/his.13922

Problems in the reproducibility of classification of small lung adenocarcinoma: an international interobserver study

Angela R Shih,[1] Hironori Uruga,[1,2] Emine Bozkurtlar,[3] Jin-Haeng Chung,[4] Lida P Hariri,[1] Yuko Minami,[5] He Wang,[6] Akihiko Yoshizawa,[7] Alona Muzikansky,[1] Andre L Moreira[8] & Mari Mino-Kenudson[1]

[1]*Massachusetts General Hospital, Boston, MA, USA,* [2]*Toranomon Hospital, Tokyo, Japan,* [3]*Marmara University, Istanbul, Turkey,* [4]*Seoul National University, Bundang Hospital, Seongnam, Republic of Korea,* [5]*National Hospital Organization, Ibarakihigashi National Hospital, Ibaraki, Japan,* [6]*Temple University School of Medicine, Philadelphia, PA, USA,* [7]*Kyoto University Hospital, Kyoto, Japan, and* [8]*NYU Langone Health, New York City, NY, USA*

Date of submission 9 February 2019
Accepted for publication 17 May 2019
Published online *Article Accepted* 20 May 2019

Shih A R, Uruga H, Bozkurtlar E, Chung J-H, Hariri L P, Minami Y, Wang H, Yoshizawa A, Muzikansky A, Moreira A L & Mino-Kenudson M

(2019) *Histopathology* **75**, 649–659. https://doi.org/10.1111/his.13922

Problems in the reproducibility of classification of small lung adenocarcinoma: an international interobserver study

Aims: The 2015 WHO classification for lung adenocarcinoma (ACA) provides criteria for adenocarcinoma *in situ* (AIS), minimally invasive adenocarcinoma (MIA) and invasive adenocarcinoma (INV), but differentiating these entities can be difficult. As our understanding of prognostic significance increases, inconsistent classification is problematic. This study assesses agreement within an international panel of lung pathologists and identifies factors contributing to inconsistent classification.

Methods and results: Sixty slides of small lung ACAs were reviewed digitally by six lung pathologists in three rounds, with consensus conferences and examination of elastic stains in round 3. The panel independently reviewed each case to assess final diagnosis, invasive component size and predominant pattern. The kappa value for AIS and MIA versus INV decreased from 0.44 (round 1) to 0.30 and 0.34 (rounds 2 and 3). Interobserver agreement for invasion (AIS versus other) decreased from 0.34 (round 1) to 0.29 and 0.29 (rounds 2 and 3). The range of the measured invasive component in a single case was up to 19.2 mm among observers. Agreement was excellent in tumours with high-grade cytology and fair with low-grade cytology.

Conclusions: Interobserver agreement in small lung ACAs was fair to moderate, and improved minimally with elastic stains. Poor agreement is primarily attributable to subjectivity in pattern recognition, but high-grade cytology increases agreement. More reliable methods to differentiate histological patterns may be necessary, including refinement of the definitions as well as recognition of other features (such as high-grade cytology) as a formal part of routine assessment.

Keywords: histological pattern, interobserver agreement, lung adenocarcinoma, minimally invasive adenocarcinoma

Introduction

Lung cancer is the leading global cause of cancer-related mortality,[1] and the high mortality rate is historically attributed to advanced stage at initial diagnosis

Address for correspondence: Mari Mino-Kenudson, Department of Pathology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA. E-mail: mminokenudson@partners.org

and limited therapeutic options for metastatic disease.[2] However, in the last two decades there has been an improvement in survival of a subset of patients with advanced non-small cell lung cancer due to the discovery of new molecular targets, histology-directed chemotherapeutic regimens and immunotherapy.[3] Additionally, lung cancer screening with low-dose computed tomography (LDCT) has been found in randomised clinical trials to improve detection of early stage lung cancers, particularly adenocarcinoma (ACA),[4–6] and has recently been implemented.[7] Thus, it has never been more critical to consistently classify early-stage lung ACA for the prediction of patient outcomes and identification of patients who may benefit from adjuvant therapy.

In an effort to create a classification system that is both biologically and clinically relevant, multidisciplinary panels from the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS) and the European Respiratory Society (ERS) put forth an updated lung ACA classification system in 2011, which recognises five histological patterns (lepidic, acinar, papillary, micropapillary and solid) and four variants (invasive mucinous, fetal lung, enteric and colloid).[8] This classification system was later adopted in the 2015 *World Health Organisation* (*WHO*) *Classification of Tumors of the Lung, Pleura, Thymus, and Heart*.[9,10] Specific criteria for diagnosis of adenocarcinoma *in-situ* (AIS), minimally invasive adenocarcinoma (MIA) and invasive adenocarcinoma (INV) have been published, with separation of AIS and MIA from INV due to the excellent prognosis of the former two diagnoses.[11–13] One of the criteria for this distinction involves an estimate of the composite percentage of the histological patterns (and variants) within a given tumour; a designation of AIS, MIA and INV is based, in part, on the proportions of lepidic and non-lepidic patterns. This classification system has been demonstrated to have improved ability to predict recurrence and prognosis, with accumulating evidence that the invasive tumour size may be of more prognostic value than the overall tumour size.[11,13–16]

As a result, the recently published *American Joint Committee on Cancer* (*AJCC*) *cancer staging manual* (8th edn) has advised that staging of lung ACA should be based in part on a measurement of the invasive component of a given tumour, separating out non-lepidic from lepidic patterns.[17] However, in routine practice, differentiating between specific histological patterns can be difficult due to poor interobserver agreement, and inconsistent recognition of these histological patterns can have a profound impact on the final diagnosis and tumour–necrosis–metastasis (TNM) stage under these new guidelines. Of note, in accordance with the implementation of lung cancer screening with LDCT, we have experienced an increasing number of resections for small, lepidic-predominant ACA, and differentiating non-lepidic from lepidic patterns can be extremely challenging in such tumours. To further evaluate this issue, this study assesses agreement within an international panel of lung pathologists and identifies factors contributing to potential problems in implementing the WHO and AJCC criteria in small lung ACA.

Materials and methods

The case files at the Massachusetts General Hospital were searched between 2006 and 2016 to identify lung ACA that had a range of diagnoses, varying from AIS to MIA to INV. Based on the extent of tumour sampling and availability of histology slides and blocks, 60 representative cases were selected for the study cohort by three authors who were not members of the observer panel (A.R.S., H.U. and M.M.K.). During the selection process, an effort was made to include cases in which it would be difficult to differentiate between AIS versus MIA and MIA versus INV on a single slide containing tumour measuring $\leq$2.5 cm (Table S1). The selected single representative block from each case was processed with consecutive sections and haematoxylin and eosin (H&E) and elastic stains. Both sections were scanned via whole slide imaging (Hamamatsu Photonics, Shizuoka, Japan), and the images were uploaded onto a server compatible for download with a web-based viewer with a built-in digital ruler. This study was approved by the institutional review board of the Massachusetts General Hospital, Boston, MA, USA.

Our observer panel consisted of six lung pathologists from three regions (three from Asia, two from the United States and one from Europe), with a range of 6–20 years of experience. Each of the 60 cases was reviewed digitally by six pathologists in three sequential rounds, with intervening time intervals between case examination (washout periods) lasting several months (Figure 1). At round 1, the panel independently reviewed each case to assess predominant pattern, invasive component size and final diagnosis (AIS, MIA or INV) as they would routinely apply to surgical cases without specific instructions to provide a baseline assessment on H&E sections. Round 2 occurred after discussion at a consensus conference, which was held as an online meeting

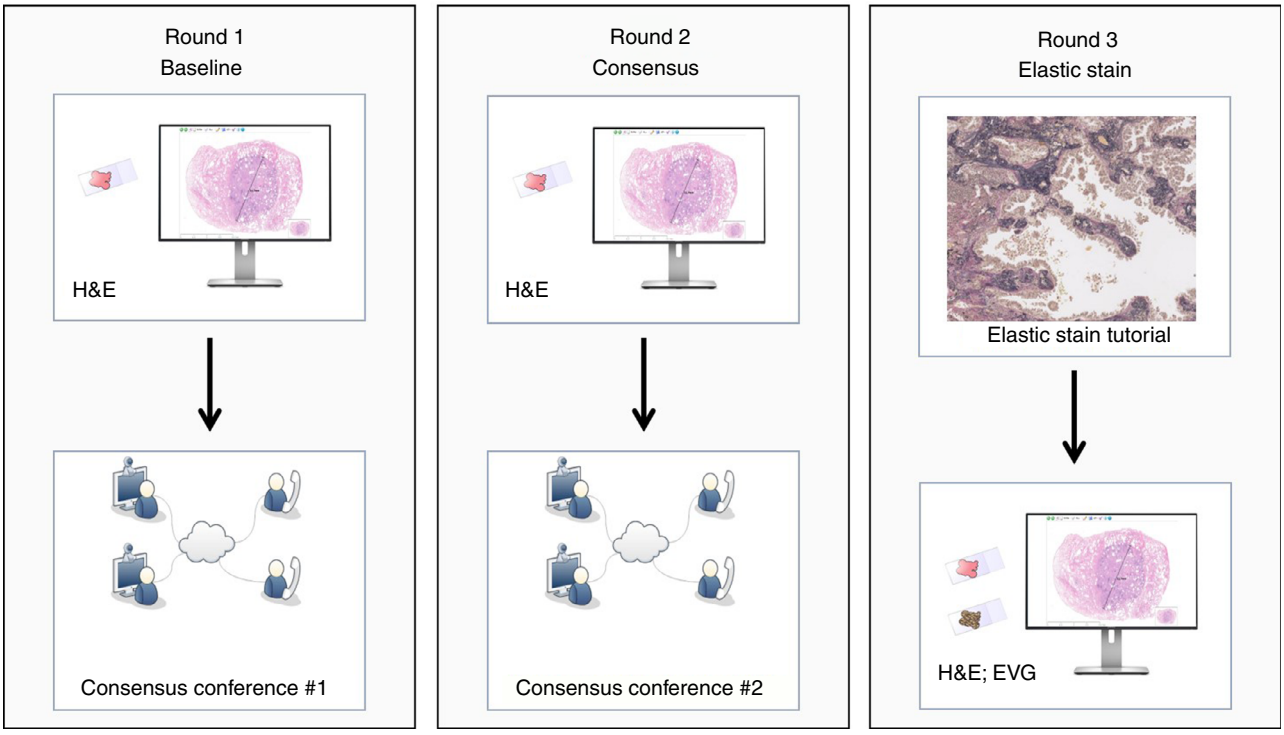

Figure 1. Study design. A single representative slide from each of 60 small lung adenocarcinoma cases were digitally scanned and reviewed by a panel of six international lung pathologists. After a consensus conference and an appropriate wash-out period of 2–3 months, the slides were re-reviewed by the same panel. After a second consensus conference, the slides were reviewed for a third time concomitantly with a consecutive elastic stain.

with review and discussion of 11 cases via selected images. The consensus conference resulted in general agreement regarding specific issues and specific cases (Data S1), and a consensus document was circulated to reiterate these points. After round 2, a tutorial was given regarding appropriate interpretation of elastic stains to highlight alveolar architecture, to facilitate differentiation of lepidic growth from other histological patterns. Subsequently, round 3 of assessment was undertaken with concomitant examination of consecutive elastic stains.

Each case was also evaluated by two authors (A.R.S. and M.M.K.) involved in case selection for nuclear grade, in which grade 1 was defined as round regular nuclei (up to ×2–3 the size of a lymphocyte) with evenly dispersed chromatin and inconspicuous nucleoli; grade 2 was defined as round, mildly irregular, minimally pleomorphic nuclei (up to ×2–3 the size of a lymphocyte) with discernible nucleoli; and grade 3 was defined as pleomorphic nuclei (typically greater than ×5 the size of a lymphocyte) with prominent nucleoli.[18,19] The tumours were further evaluated for cytological grade, in which low grade was defined as a low degree of cellular pleomorphism with a small cell size as categorized by expert pathologists, and high grade was defined as a high degree of cellular pleomorphism with a large cell size.

Statistical analysis was performed for each round of evaluation in the form of Fleiss' kappa coefficient for multirater agreement. Analysis was performed in three score groups. Score 0 compared each diagnosis separately (AIS versus MIA versus INV); score 1 compared the raters' abilities to separate out non-invasive ACA (AIS versus MIA and INV); and score 2 compared the raters' abilities to separate out ACA with favourable prognosis (AIS and MIA versus INV). Interobserver agreement in the predominant histological pattern of growth identified by the observers was also calculated by comparing each histological pattern separately (lepidic, acinar, papillary, solid, and micropapillary). Further analysis was also done to assess the agreement among the observers in identifying predominantly lepidic growth compared to all other invasive patterns. Intraobserver concordance in final diagnosis across the rounds was also assessed. Statistical analysis was performed using SAS version 9.4 software and the %MAGREE macro,[20] as well as online interobserver agreement calculators.[21]

Results

INTER- AND INTRAOBSERVER AGREEMENTS ON FINAL DIAGNOSIS

The overall Fleiss' kappa coefficient for score 0 (AIS versus MIA versus INV) decreased from round 1 ($\kappa = 0.31$) to round 2 ($\kappa = 0.23$), with no improvement from round 2 to round 3 ($\kappa = 0.24$); in general, the data indicate fair agreement in all rounds. Similarly, the kappa coefficient for score 1 (AIS versus MIA and INV) also decreased between rounds from 0.34 (round 1) to 0.29 (round 2) and 0.29 (round 3). The kappa value for score 2 (AIS and MIA versus INV) also decreased between rounds from round 1 ($\kappa = 0.44$; moderate agreement) to round 2 ($\kappa = 0.30$; fair agreement) and round 3 ($\kappa = 0.34$; fair agreement) (Figure 2A). Interestingly, in round 1, the raters had 100% agreement on final diagnosis in 12 cases (AIS, $n = 3$; MIA, $n = 4$; INV, $n = 5$), which decreased in round 3 to only six cases (AIS, $n = 1$; MIA, $n = 1$; INV, $n = 4$).

The intraobserver agreement between rounds 1 and 2 ranged from 0.57 to 0.78, while the intraobserver agreement between rounds 2 and 3 ranged from 0.62 to 0.80. Between rounds 1 and 3, intraobserver agreement ranged from 0.53 to 0.73. In general, the more junior pathologists (<10 years of experience) showed decreased intraobserver agreement compared to the more senior pathologists. Overall, however, participants predominantly maintained a stable intraobserver

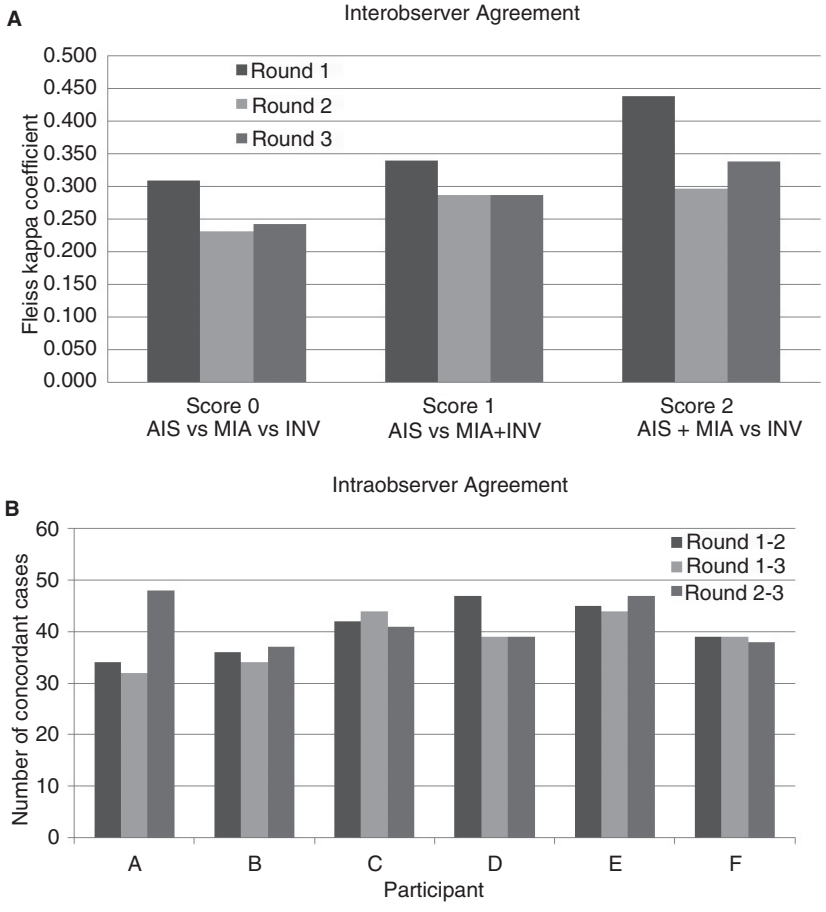

Figure 2. Inter- and intraobserver agreements in overall diagnosis. **A**, Interobserver reliability in overall diagnosis was assessed by Fleiss' kappa coefficient, which decreased for score 0 [adenocarcinoma *in situ* (AIS) versus minimally invasive adenocarcinoma (MIA) versus invasive adenocarcinoma (INV)] from round 1 ($\kappa = 0.309$) to round 2 ($\kappa = 0.232$), with a minimal improvement from rounds 2 to 3 ($\kappa = 0.243$). The kappa coefficient for score 1 (AIS versus MIA and INV) also decreased between rounds from 0.340 (round 1) to 0.286 (round 2) and 0.287 (round 3). The kappa value for score 2 (AIS and MIA versus INV) also decreased between rounds from round 1 ($\kappa = 0.438$; moderate agreement) to round 2 ($\kappa = 0.296$; fair agreement), with a subsequent increase in round 3 ($\kappa = 0.338$; fair agreement). **B**, Intraobserver agreement between rounds 1 and 2 ranged from 0.57 to 0.78, while the concordance between rounds 2 and 3 ranged from 0.62 to 0.80. Between rounds 1 and 3, intraobserver concordance ranged from 0.53 to 0.73. Overall, participants predominantly maintained a relatively stable intraobserver concordance between rounds, with concordant diagnoses in an average of 40 of the total 60 cases evaluated.

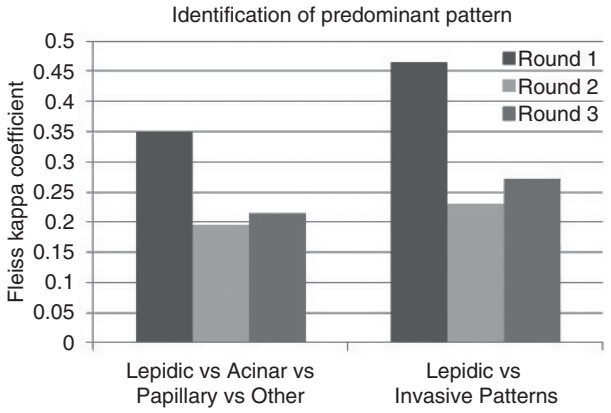

Figure 3. Interobserver agreement in identification of predominant histological pattern. The interobserver agreement in the predominant histological pattern of growth, as calculated by Fleiss' kappa coefficient, ranged from fair to moderate across all rounds in comparing each histological pattern separately as well as in comparing lepidic growth from all invasive patterns. As with agreement on final diagnosis, the kappa coefficient showed fair to moderate agreement in round 1 (0.35 and 0.47, respectively), but decreased in round 2 (0.20 and 0.23, respectively) and round 3 (0.21 and 0.27, respectively).

agreement between rounds, with concordant diagnoses in an average of 40 of the total 60 cases evaluated (Figure 2B). Interestingly, a comparison of interobserver agreement in the 19 cases with high levels of intraobserver concordance across the raters (80–100% concordance between rounds 1 and 3) showed predominantly moderate interobserver agreement in these cases, indicating that there was better agreement in presumably straightforward cases. In the eight cases with low levels of intraobserver concordance (0–40% concordance between rounds 1 and 3) there was predominantly poor agreement across rounds, with findings that suggest significant inconsistency in the observers' abilities to distinguish MIA from INV (Figure S1).

Misinterpretation of the WHO criteria for MIA resulted in 18 instances of misclassification across all raters in round 1, which decreased to four instances in round 3 after multiple consensus conferences. Using the WHO criteria as the gold standard, the most common issue by far was incorrect classification as MIA in cases with a predominant pattern that was not evaluated as lepidic by the observer. Additionally, a mucinous ACA case had a wide range of diagnosis among raters in all rounds, ranging from AIS to INV.

CLASSIFICATION OF PREDOMINANT PATTERN

The most frequent predominant consensus pattern (that a majority of the observers diagnosed in each case, using the assessment of the authors who selected the cases as a 'tiebreaker' if necessary) was lepidic (75, 75 and 85%), followed by papillary (13, 8 and 8%), acinar (7, 10 and 2%), mucinous adenocarcinoma (2, 2 and 2%) and micropapillary (2, 0 and 3%) in rounds 1, 2 and 3, respectively. No consensus was reached in one case in round 1 and three cases in round 2; consensus was reached in all cases in round 3. The interobserver agreement in the predominant pattern, as calculated by Fleiss' kappa coefficient, ranged from fair to moderate across all rounds in comparing each histological pattern separately as well as in comparing lepidic growth to all invasive patterns (Figure 3). As with agreement on final diagnosis, the kappa coefficient showed fair to moderate agreement in round 1 (0.35 and 0.47, respectively), but decreased to fair agreement in round 2 (0.20 and 0.23, respectively) and round 3 (0.21 and 0.27, respectively).

MEASUREMENT OF INVASIVE SIZE

The range of the measurement of the invasive component for each case is presented for rounds 1, 2 and 3 (Figure 4). Comparison of the three graphs shows that the general trend of the mean measured invasive component remained relatively similar for the six observers. However, there was a marked expansion of the range across raters for most cases in rounds 2 and 3. For instance, in round 1, the largest range in measured invasion in a single case was 0–13 mm; in round 2 it was 0.8–20 mm in a different case; and in round 3 it was 0.9–16 mm in a third case.

EFFECTS OF NUCLEAR AND CYTOLOGICAL GRADES ON INTEROBSERVER AGREEMENT

The cases were also rated for nuclear morphology by two authors involved in the case selection (A.R.S. and M.M.K.), and each tumour was classified as nuclear grades 1 ($n = 7$), 2 ($n = 47$) or 3 ($n = 6$). Agreement was moderate for the grade 2 tumours (Fleiss' kappa ranging from 0.19 to 0.35; Figure S2). The observers did not agree on the diagnosis of any grade 1 tumours ($n = 7$) in rounds 1–2 and only universally agreed in one case of AIS in round 3. The six observers made a diagnosis of INV on grade 3 tumours ($n = 6$) in 30 (83%), 28 (78%) and 29 (81%) of 36 total diagnoses each in rounds 1, 2 and 3, respectively, while no observer called AIS on any of the grade 3 tumours across all rounds. However, only limited numbers of cases of nuclear grades 1 and 3 were present in this study and Fleiss' kappa coefficient could not be calculated for these groups.

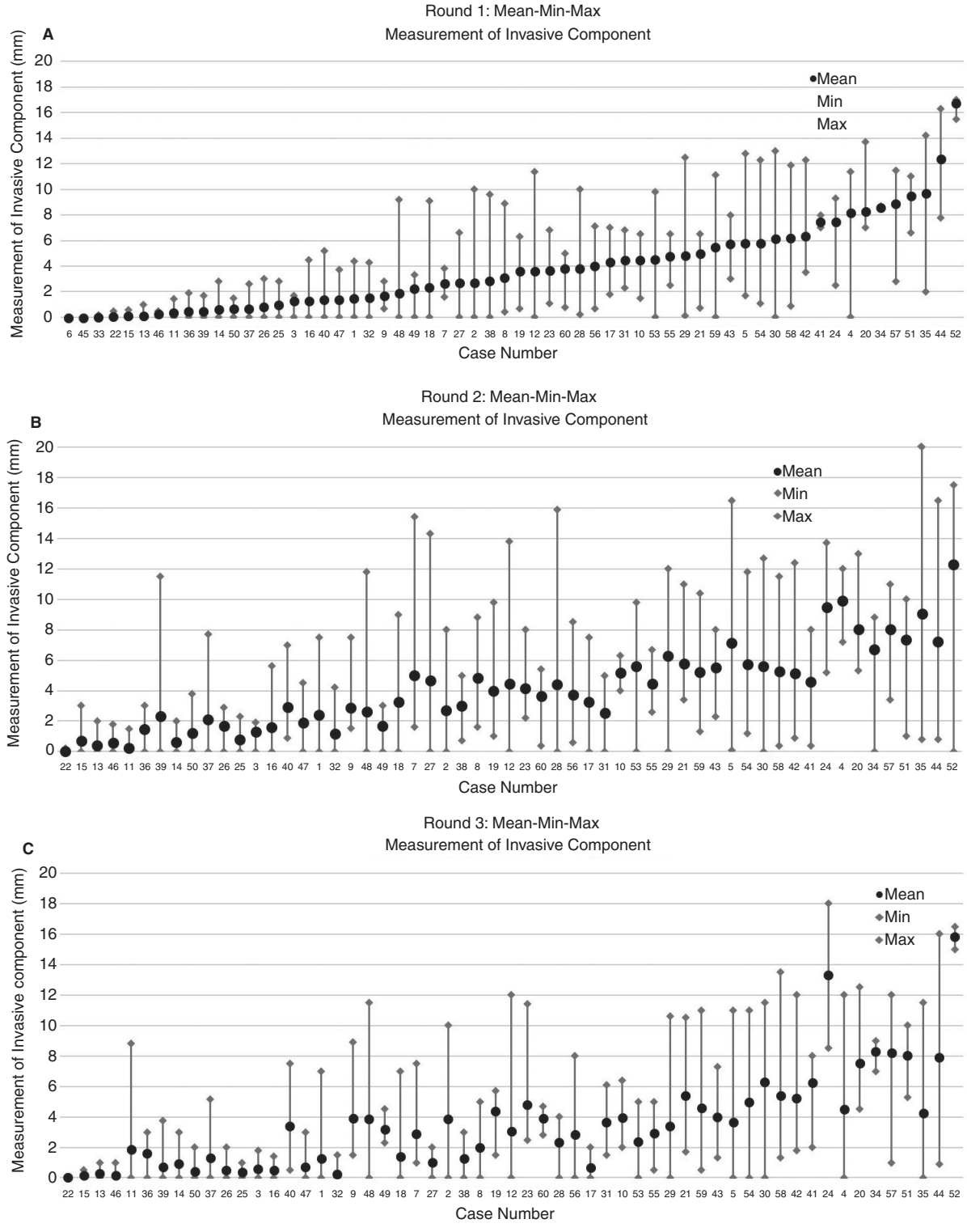

Figure 4. Variability in measurement of invasive component across observers. The mean and range of the measured invasive component for each case is presented for round 1 (**A**), round 2 (**B**) and round 3 (**C**). The black circle represents the mean and the grey diamonds and bar represent the range. Comparison of the three rounds shows mild variability with retention of the overall trend across the cases in rounds 2 and 3; however, the range increased in rounds 2–3 in comparison to the baseline assessment in round 1.

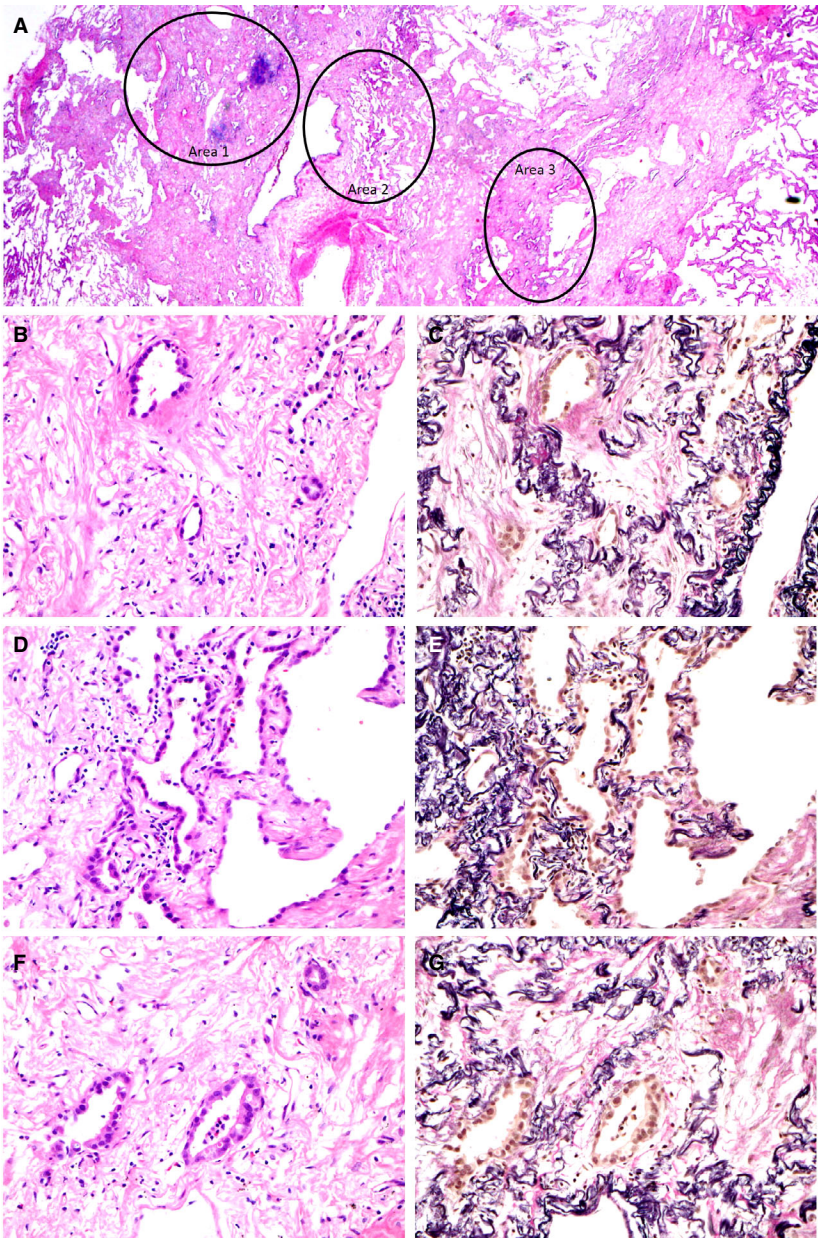

Figure 5. Differentiating lepidic and acinar patterns in multiple foci in a background of fibroelastosis. To illustrate difficulties in identification of small infiltrating glands, **A** shows a low magnification image of an adenocarcinoma with multiple foci of possible acinar pattern around the periphery of the fibroelastosis. Area 1 (**B**) shows focal disruption of an elastic framework by a gland associated with collagen deposition (**C**, elastic stain); area 2 (**D**) shows lepidic growth with a preserved elastic framework (**E**, elastic stain); and area 3 (**F**) shows disruption of an elastic framework by glands and collagen deposition (**G**, elastic stain). Although no fibroblastic proliferation or classic desmoplastic reaction is seen, the majority of observers classified areas 1 and 3 as acinar pattern.

The cases were further classified by the two authors as cytologically low grade (n = 57) and high grade (n = 3). Again, while the number of cytologically high-grade cases was limited, there was excellent agreement for the cytologically high-grade tumours. The six observers made a diagnosis of INV on cytologically high-grade tumours (n = 3) in 18 (100%), 15 (83%) and 17 of 18 (94%) total diagnoses each in rounds 1, 2 and 3, respectively, while no observer called AIS on any of the cases across all rounds. However, Fleiss' kappa coefficient could not be calculated for this group due to the small sample size. There was only fair agreement for the cytologically low-grade tumours (Fleiss' kappa ranging from 0.21 to 0.38; Figure S3).

Discussion

Interobserver agreement on the diagnosis of small lung ACAs between six observers was fair to moderate, and did not improve substantially after consensus conferences and with concomitant elastic stain evaluation. As prior studies have shown no survival or therapeutic difference between AIS and MIA, the kappa value of greatest significance is that of score 2, which measures the ability of the observers to

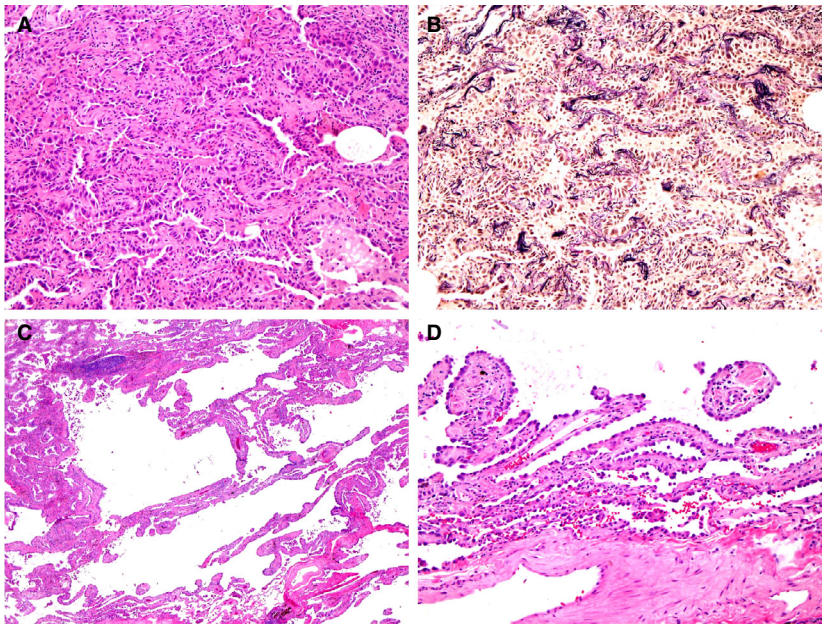

Figure 6. Differentiating lepidic, acinar and papillary patterns. Representative images of two examples illustrating problematic issues in histological pattern classification are shown here. A high magnification image (**A**) shows an area of alveolar collapse with foci suggestive of an acinar-type growth pattern, but an elastic-Van Gieson (EVG) stain (**B**) highlights a preserved elastic framework in the collapsed lung parenchyma, confirming the entire area to be lepidic. A second difficult scenario is in emphysematous areas, where disruption of the alveolar framework with mild interstitial thickening can mimic papillary growth (**C**,**D**).

separate out INV. In this study, however, the interobserver agreement for score 2 was predominantly found to be fair, which may be clinically problematic. We found that subjectivity in the diagnosis of lepidic pattern was primarily responsible for the lack of agreement, as discordant interpretations of lepidic and other patterns led to substantial variation in measurement of the invasive component and final diagnosis. The difference in measurement of the invasive component among raters in a single case was up to 19.2 mm, which is well beyond the 5 mm measurement minimum requirement for a diagnosis of INV. Initially, misinterpretation of the WHO criteria contributed to inconsistent classification, but these instances decreased substantially with education. However, the fact that there was poor interobserver agreement and relatively stable intraobserver agreement suggests that the findings in this study are due to real and reproducible differences in opinion.

Several issues were recognised as impacting classification of lung ACA under the new guidelines and hindering better agreement among pathologists. These issues included histological pattern recognition, classification of mucinous tumours and measurement of multifocal invasion, and are described in depth in Data S1. Of these issues, disagreement on histological pattern recognition was the most problematic, particularly confidently differentiating between lepidic and well-differentiated acinar or papillary patterns. The difficulty in differentiating lepidic from acinar patterns could be attributed to histological identification of stromal invasion (Figure 5) and parenchymal collapse (Figure 6A,B), while disruption of the background lung architecture secondary to emphysema probably contributes to the inconsistent differentiation of lepidic from papillary patterns (Figure 6C,D). Of note, for the histological identification of stromal invasion, some observers emphasised the importance of seeing fibroblastic proliferation or desmoplasia, with abnormal growth disrupting normal alveolar structures. However, the majority noted that invasion can occur without definitive histological evidence of desmoplasia and may be based solely on the recognition of specific histological patterns.

Examination of elastic stains can be helpful in certain situations by highlighting the underlying alveolar architecture. For instance, the disruption of the underlying elastic framework by glands suggests invasion (i.e. acinar pattern), even in very small foci (Figure 5A–C, F–G), while an intact elastic framework confirms lepidic pattern (Figure 5D,E). Similarly, an elastic stain can reliably highlight the intact alveolar framework in lepidic pattern involved in parenchymal collapse (Figure 6A,B). In general, at least some Japanese pathologists have

been using the elastic stain as a method of outlining the underlying alveolar architecture for much longer than western pathologists,[22] and consequently have more experience in using it to distinguish purely lepidic growth from true invasion. Because non-Japanese pathologists may be less familiar with interpretation of the elastic stain, a tutorial was provided before round 3 of evaluation. Unfortunately, this did not appear to substantially improve agreement, and it is unclear whether further discussion on the appropriate interpretation of these stains would have been helpful. It is also possible that digital evaluation of concomitant H&E and elastic stains was extremely difficult due to the inconvenience in manoeuvering the slides to the identical field of interest and the inability to visualise both stains simultaneously with the virtual imaging system that we employed.

The interobserver agreement on classification of lung ACAs identified in this study is poor and somewhat worse than those found in previous reproducibility studies. In a study by Thunnissen *et al.*[23] involving evaluation of a representative microphotograph of lung ACAs by a panel of 26 expert lung pathologists, classic cases were found to have substantial reproducibility, with a kappa value of 0.77, but more difficult cases (such as those used in this study) had much poorer agreement, with a kappa value of 0.38. Similarly, a reproducibility study for identification of a predominant pattern in randomly selected, non-mucinous lung ACAs evaluated by participants in their entirety was found to have moderate to substantial agreement among lung pathologists (kappa value ranging from 0.44 to 0.72) and poor to moderate agreement among non-lung pathologists (kappa value ranging from 0.38 to 0.47); however, agreement among non-lung pathologists improved substantially with training, suggesting that when evaluation is performed by specifically trained pathologists interobserver variability is acceptable.[24]

There are multiple possible explanations as to why this study's interobserver agreement is substantially poorer than prior studies. First, the cases used in this study were small tumours, the vast majority of which exhibited lepidic predominant growth pattern; in contrast, in the study by Warth and colleagues, the most frequent predominant consensus pattern was solid (37%) followed by acinar (35%).[24] Further, most of our study cohort showed low-grade cytology, potentially enhancing the difficult in differentiating lepidic from papillary or acinar patterns. The issue was elucidated by Figure 4B,C of Thunnissen *et al.*, in which cases exhibiting high-grade cytology were unanimously diagnosed as invasive, while cases exhibiting low-grade cytology but otherwise similar morphology resulted in split opinions.[23] In our study, when stratified by nuclear grade and cytologic grade, agreement among the observers stratifies similarly, with good agreement in cases with high grade nuclear and cytologic grade and a range of opinions in cases with low nuclear and cytologic grade. . Although the numbers are limited in this study, interobserver agreement was significantly higher in cases with high nuclear and high cytological grades. These cases showed high cytological features in non-lepidic patterns and were almost universally classified as INV by the observers. This can be explained either by a biological association of high-grade cytology with invasive architectural patterns, or by a tendency of pathologists to associate nuclear atypia and cytological pleomorphism with more aggressive biological behaviour. In either case, evaluation of cytological features as a component of classification may help to improve agreement.

A second possible reason for the poor interobserver agreement in this study is that our panel was asked to render a diagnosis based on a single digital slide. It is possible that our agreement on predominant histological pattern and classification might have been better if the observers had evaluated the entire tumour with multiple slides; however, Warth and colleagues have shown improved interobserver agreement in cases with evaluation of one to three physical slides compared to evaluation of four to 12 slides.[24] Additionally, the impact of digital evaluation and digital measurement of invasion as opposed to conventional evaluation with a physical slide at a microscope cannot be completely excluded. However, a variety of small studies have suggested a high level of diagnostic concordance between whole slide imaging and light microscopy, although further study is necessary to be confident in these findings.[25]

Lastly, the multiple rounds of evaluation with the emphasis on improving interobserver agreement may have resulted in pathologist indecision rather than clarification. Asking junior pathologists to re-evaluate their criteria for specific histological types can result in hesitation and over-deliberation that may have contributed to the expansion of the range of the measured invasive component in rounds 2 and 3. In contrast, we understand that it can be difficult for senior and expert lung pathologists to re-evaluate and alter their long-held criteria for recognition of specific histological types. Thus, there is a need to clarify issues of histological evaluation that are causing these differences of opinion and define each pattern in a more instructive manner.

Most importantly, discussion among the participants indicated a strong belief that distinguishing

between AIS and MIA and between MIA and AIS based on small foci suspicious for invasion may not be critical if the overarching pattern is lepidic in small lung ACAs.[26] Although the differentiation of AIS versus MIA versus INV is still important from the perspective of staging, the biological potential of these tumours is similar; consequently, discriminating between the three in lepidic-predominant small tumours may not be essential from a patient treatment stand-point. Similarly, distinguishing between invasive patterns may not be critical for overall diagnosis if the overarching pattern is invasive. However, reliably identifying more aggressive solid and micropapillary components is still important because of their association with a worse prognosis; even small (5%) components of micropapillary pattern have been associated with unfavourable patient outcomes.[27–31] Further, adjuvant chemotherapy is reportedly more effective in ACA with a predominant solid or micropapillary pattern.[32] More work should be conducted aimed at improving consistency in identifying these patterns with a poor prognosis.

What this study demonstrates is that poor agreement, even among expert lung pathologists, leads to inconsistent classification of tumours under the current WHO and AJCC staging guidelines. The heterogeneity of lung ACA gives rise to several scenarios in which the practical application of the guidelines will be difficult. The question of how we can improve agreement in histological pattern recognition (particularly the differentiation of lepidic from non-lepidic patterns) is a difficult one without any clear answers. While an elastic stain was not demonstrated to be particularly helpful in this study, it is possible that further education on the interpretation of these stains may enhance their utility. Further studies involving large numbers of international experts examining small lung adenocarcinomas may identify more specific issues associated with histological pattern recognition, and subsequently would clarify the definition of lepidic pattern as well as other recognized histologic patterns in a more instructive manner. The finding that high-grade nuclei/cytology increases agreement raises the question of whether it may be helpful as a formally recognised component of classification, given that high-grade nuclei/cytology is a predictor of poor prognosis.[18,19,33–35]

In summary, consistent classification of lung ACA, particularly small tumours, is challenging, due largely to substantial differences in histological pattern recognition among pathologists, while the new guidelines as outlined for diagnosis by the WHO and for staging by the AJCC have made histological pattern recognition, in particular the differentiation of lepidic versus non-lepidic patterns, critically important. More reliable methods to differentiate histological patterns, including refinement of the definition of each pattern as well as education, are necessary to improve interobserver agreement to allow for consistent classification of tumours to provide appropriate care to lung cancer patients.

Acknowledgements

We would like to acknowledge Yukako Yagi PhD, Xiujun Fu PhD and Pinky Bautista PhD for their assistance with whole slide imaging. We would also like to acknowledge Stephen Conley for his assistance in organising the online consensus conferences. There are no sources of funding for this study.

Conflict of Interest

None.

References

1. Jemal A, Ward EM, Johnson CJ *et al*. Annual report to the nation on the status of cancer, 1975–2014, featuring survival. *J. Natl Cancer Inst.* 2017; **109**; djx030.
2. Chen VW, Ruiz BA, Hsieh M *et al*. Analysis of stage and clinical/prognostic factors for lung cancer from SEER registries: AJCC staging and collaborative stage data collection system. *Cancer*. 2014; **120**; 3781–3792.
3. Ferrara R, Mezquita L, Besse B. Progress in the management of advanced thoracic malignancies in 2017. *J. Thorac. Oncol.* 2018; **13**; 301–322.
4. Aberle DR, Adams AM, Berg CD *et al*. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J Med.* 2011; **365**; 395–409.
5. Fintelmann FJ, Bernheim A, Digumarthy SR *et al*. The 10 pillars of lung cancer screening: rationale and logistics of a lung cancer screening program. *Radiographics* 2015; **35**; 1893–1908.
6. Moyer VA, Force PU. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Ann. Intern. Med.* 2014; **160**; 330–338.
7. Jaklitsch MT, Jacobson FL, Austin J *et al*. The American Association for Thoracic Surgery guidelines for lung cancer screening using low-dose computed tomography scans for lung cancer survivors and other high-risk groups. *J. Thorac. Cardiovasc. Surg.* 2012; **144**; 33–38.
8. 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**; 244–285.
9. Travis WD, Brambilla E, Nicholson AG *et al*. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. *J. Thorac. Oncol.* 2015; **10**; 1243–1260.

10. Travis WD, Brambilla E, Burke AP *et al*. *WHO classification of Tumours of the Lung, Pleura, Thymus and Heart (World Health Organization classification of tumours)*. 4th ed. Lyon: International Agency for Research on Cancer, 2015.
11. Warth A, Muley T, Meister M *et al*. The Novel Histologic International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society Classification System of Lung Adenocarcinoma is a stage-independent predictor of survival. *J. Clin. Oncol.* 2012; **30**; 1438–1446.
12. Yoshizawa A, Sumiyoshi S, Sonobe M *et al*. Validation of the IASLC/ATS/ERS lung adenocarcinoma classification for prognosis and association with EGFR and KRAS gene mutations: analysis of 440 Japanese patients. *J. Thorac. Oncol.* 2013; **8**; 52–61.
13. 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**; 653.
14. Russell PA, Wainer Z, Wright GM *et al*. Does lung adenocarcinoma subtype predict patient survival? A clinicopathologic study based on the new International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary lung adenocarcinoma classification. *J. Thorac. Oncol.* 2011; **6**; 1496–1504.
15. Yanagawa N, Shiono S, Abiko M *et al*. New IASLC/ATS/ERS classification and invasive tumor size are predictive of disease recurrence in stage I lung adenocarcinoma. *J. Thorac. Oncol.* 2013; **8**; 612–618.
16. Kameda K, Eguchi T, Lu S *et al*. Implications of the eighth edition of the TNM proposal: invasive versus total tumor size for the T descriptor in pathologic stage I-IIA lung adenocarcinoma. *J. Thorac. Oncol.* 2018; **13**; 1919–1929.
17. Amin MB, Edge S, Greene F *et al*. *AJCC cancer staging manual*. Vol. xvii. 8th ed. New York, NY: Springer International Publishing, 2017.
18. von der Thüsen JH, Tham YS, Pattenden H *et al*. Prognostic significance of predominant histologic pattern and nuclear grade in resected adenocarcinoma of the lung: potential parameters for a grading system. *J. Thorac. Oncol.* 2013; **8**; 37–44.
19. Barletta JA, Yeap BY, Chirieac LR. Prognostic significance of grading in lung adenocarcinoma. *Cancer* 2010; **116**; 659–669.
20. Fleiss JL. *Statistical methods for rates and proportions*. Chichester: John Wiley & Sons Inc, 2003.
21. Geertzen J. Inter-rater agreement with multiple raters and variables. Available at: https://mlnl.net/jg/software/ira/ (accessed February 2019).
22. Sakurai H, Maeshima A, Watanabe S *et al*. Grade of stromal invasion in small adenocarcinoma of the lung: histopathological minimal invasion and prognosis. *Am. J. Surg. Pathol.* 2004; **28**; 198.
23. Thunnissen E, Beasley M, Borczuk AC *et al*. Reproducibility of histopathological subtypes and invasion in pulmonary adenocarcinoma. An international interobserver study. *Mod. Pathol.* 2012; **25**; 1574.
24. Warth A, Stenzinger A, von Brünneck A-C *et al*. Interobserver variability in the application of the novel IASLC/ATS/ERS classification for pulmonary adenocarcinomas. *Eur Respir. J.* 2012; **40**; 1221–1227.
25. Goacher E, Randell R, Williams B *et al*. The diagnostic concordance of whole slide imaging and light microscopy: a systematic review. *Arch. Pathol. Lab. Med.* 2016; **141**; 151–161.
26. Kadota K, Villena-Vargas J, Yoshizawa A *et al*. Prognostic significance of adenocarcinoma in situ, minimally invasive adenocarcinoma, and nonmucinous lepidic predominant invasive adenocarcinoma of the lung in patients with stage I disease. *Am. J. Surg. Pathol.* 2014; **38**; 448–460.
27. Hung J-J, Jeng W-J, Chou T-Y *et al*. Prognostic value of the new International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society lung adenocarcinoma classification on death and recurrence in completely resected stage I lung adenocarcinoma. *Ann. Surg.* 2013; **258**; 1079–1086.
28. Ohe M, Yokose T, Sakuma Y *et al*. Stromal micropapillary component as a novel unfavorable prognostic factor of lung adenocarcinoma. *Diagn. Pathol.* 2012; **7**; 3.
29. Ujiie H, Kadota K, Chaft JE *et al*. Solid predominant histologic subtype in resected stage I lung adenocarcinoma is an independent predictor of early, extrathoracic, multisite recurrence and of poor postrecurrence survival. *J. Clin. Oncol.* 2015; **33**; 2877–2884.
30. Zhang Y, Li J, Wang R *et al*. The prognostic and predictive value of solid subtype in invasive lung adenocarcinoma. *Sci. Rep.* 2015; **4**; srep07163.
31. Zhao Y, Wang R, Shen X *et al*. Minor components of micropapillary and solid subtypes in lung adenocarcinoma are predictors of lymph node metastasis and poor prognosis. *Ann. Surg. Oncol.* 2016; **23**; 2099–2105.
32. Tsao M-S, Marguet S, Teuff G *et al*. Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. *J. Clin. Oncol.* 2015; **33**; 3439–3446.
33. Mäkinen JM, Laitakari K, Johnson S *et al*. Histological features of malignancy correlate with growth patterns and patient outcome in lung adenocarcinoma. *Histopathology* 2017; **71**; 425–436.
34. Asamura H, Ando M, Matsuno Y *et al*. Histopathologic prognostic factors in resected adenocarcinomas is nuclear DNA content prognostic? *Chest* 1999; **115**; 1018–1024.
35. Nakazato Y, Minami Y, Kobayashi H *et al*. Nuclear grading of primary pulmonary adenocarcinomas. *Cancer* 2010; **116**; 2011–2019.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Comparison of interobserver agreement in cases with high and low intraobserver concordance.

Figure S2. Interobserver agreement stratified by nuclear grade.

Figure S3. Interobserver agreement stratified by cytologic grade.

Figure S4. Classification of an unusual morphologic patterns.

Table S1. Tumor cohort characteristics and diagnoses.

Data S1. Consensus conference discussion points: problematic issues.