Eur Respir J 2012; 40: 1221–1227
DOI: 10.1183/09031936.00219211

Interobserver variability in the application of the novel IASLC/ATS/ERS classification for pulmonary adenocarcinomas

Arne Warth*, Albrecht Stenzinger*, Ann-Christin von Brünneck#, Benjamin Goeppert*, Judith Cortis*, Iver Petersen¶, Hans Hoffmann+, Philipp A. Schnabel* and Wilko Weichert*

ABSTRACT: Recently, a novel classification for pulmonary adenocarcinomas (ADCs) was published, the cornerstone of which is the quantification of growth patterns. Data on reproducibility in the routine diagnostic setting are lacking.

100 constitutive cases of lung ADC resection specimens from our archives were classified independently by five pulmonary pathologists and two residents according to the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society classification.

The most frequent predominant pattern in our cohort was solid (37%), followed by acinar (35%), lepidic (20%), papillary (5%) and micropapillary (3%). κ-values for the denomination of the predominant pattern revealed substantial agreement for pulmonary pathologists (κ=0.44–0.72) and fair agreement for residents (κ=0.38–0.47). Interobserver variability was significantly higher in cases with higher slide numbers (p=0.028) and was considerably reduced after training. Intraobserver variability was low (κ=0.79–0.87). Papillary and micropapillary patterns were the most complicated patterns to evaluate, while evaluation of lepidic and solid tumour growth was straightforward.

Our data imply that the novel classification of pulmonary ADC is applicable with acceptable interobserver variability if performed by specifically trained pathologists. Additional efforts are needed to harmonise the application of this novel and clinically important classification scheme of pulmonary ADC.

KEYWORDS: IASLC/ATS/ERS classification, interobserver agreement, pulmonary adenocarcinoma, reproducibility

AFFILIATIONS
*Institute of Pathology, University Hospital Heidelberg,
+Dept of Thoracic Surgery, Thoraxklinik Heidelberg, Heidelberg,
#Institute of Pathology, Charité University Hospital Berlin, Berlin, and
¶Institute of Pathology, University Hospital Jena, Jena, Germany.

CORRESPONDENCE
A. Warth
Institute of Pathology
University Hospital Heidelberg
Im Neuenheimer Feld 220/221
D-69120 Heidelberg
Germany
E-mail: arne.warth@med.uni-heidelberg.de

Received:
Dec 13 2011
Accepted after revision:
Feb 13 2012
First published online:
March 09 2012

Pulmonary carcinomas are very frequent neoplasms, and the leading cause of cancer deaths for both sexes in the western world [1]. Of these, the most prevalent tumour type is adenocarcinoma, accounting for approximately half of all lung cancers.

In pulmonary tumours, as well as in all other neoplasms, a reliable estimation of patient survival probability is crucial, since the selection of conventional therapy regimens administered in the palliative or adjuvant setting is largely based on the risk that the patient will die of his disease in a given time-frame. The cornerstone of clinical survival estimation in lung neoplastic disease is stage categorisation. Recently, the staging system of lung neoplasms has been revised and extended to include additional pulmonary tumour entities now comprising squamous cell carcinoma (SCC), adenocarcinoma (ADC) and pulmonary carcinoids, as well as several rare types of non-small cell lung cancer [2]. In addition, molecular alterations that may predict response to certain drugs have been introduced into clinical decision making. This comprises factors like *EGFR* mutations (for erlotinib, gefitinib) or *EML4–ALK* translocations [3–5]. Additional clinically applicable prognosticators or response predictors for pulmonary SCC and ADC have not been established. A myriad other suggested prognostic and predictive markers have not made it into the clinic yet [6] for practical, financial, technical or scientific reasons.

Tumour grade as determined by the pathologist on the basis of conventional histology is a reliable predictor of survival in a broad variety of solid human tumour entities. In some of these entities, such as breast cancer, prostate cancer or sarcoma,

European Respiratory Journal
Print ISSN 0903-1936
Online ISSN 1399-3003

standardised grading strongly influences clinical decision making [7–9].

In contrast, in the last decades, grading in lung cancer has largely been performed without internationally accepted criteria and without any clinical impact. However, an increasing number of groups is working on this issue [5, 10]. Recently, on the basis of some novel data obtained by these groups, an international multidisciplinary expert panel of the International Association for the Study of Lung Cancer (IASLC), the American Thoracic Society (ATS), and the European Respiratory Society (ERS) has proposed a novel classification system for ADC, which is based on the evaluation of tumour architecture and is intended to allow for standardised tumour grading [10]. The core of this novel classification is a categorisation of pulmonary ADCs according to their predominant growth pattern. The independent prognostic impact of this classification across all tumour stages has subsequently been confirmed by us and others in large ADC cohorts from Europe, Australia and the USA [11–14]. Thus, grading of pulmonary ADC on the basis of the predominant pattern will very probably enter clinical decision making in the near future. However, prior to the application of this classification in the daily clinical routine, it is mandatory to corroborate that this classification can be reliably and reproducibly utilised under real-life conditions by different evaluators beyond a well-defined study setting. Data on that issue, however, are currently lacking. Therefore, in the present study, we assessed intra- and interobserver variability in pulmonary ADC pattern recognition under routine diagnostic conditions. To do so, we randomly selected 100 ADC cases from our archives and performed a comprehensive round-robin test with seven pathologists with different levels of lung pathology training from three German institutions.

MATERIAL AND METHODS

Tissue

100 cases of randomly selected pulmonary ADC were evaluated. All tumours were surgically removed at the Thoracic Hospital of the University of Heidelberg (Heidelberg, Germany) in 2008. Of the respective cases, all slides processed for routine diagnosis were reviewed by all participants. To focus on the pattern issue, specific subtypes of pulmonary ADC, *e.g.* invasive mucinous ADC, were excluded.

Reviewers

A total of seven reviewers took part in the study. All were provided with the novel IASLC/ATS/ERS classification [10] without further discussion. All reviewers were blinded to the results of the other reviewers. Three of the reviewers (A. Warth, W. Weichert and P.A. Schnabel) were pathologists from the University Hospital in Heidelberg with ample expertise in pulmonary pathology ($\geqslant$5 yrs of continuous diagnostic work in this field). In addition, two expert pulmonary pathologists from other German high-volume centres were included (A-C. von Brünneck, Charité University Hospital, Berlin; I. Petersen, University Hospital Jena). The remaining two evaluators (B. Goeppert and A. Stenzinger) were pathologists in training with a certain expertise in general surgical pathology but with no special training in pulmonary pathology. Two of the evaluators (A. Warth and W. Weichert) assessed the whole slide set a second time to probe for intraobserver variability. The time elapsed between both evaluations was >1 month.

After the first round of evaluations, the two non-pulmonary pathologists (B. Goeppert and A. Stenzinger) took part in a training session during which both attended the re-evaluation of the whole slide sets of all 100 cases by either A. Warth or W. Weichert. After 1 week, both pathologists in training undertook a second solitary independent evaluation round.

Evaluation

All haematoxylin and eosin slides sampled from the primary tumours were evaluated. Lymph node or distant metastases were not included. Tumour architecture was classified as lepidic, acinar, papillary, micropapillary or solid according to the respective guidelines of the IASLC/ATS/ERS (fig. 1) [10]. Quantification was performed in 5% increments as recommended. Tumours were categorised according to the pattern with the highest percentage. In addition, a consensus categorisation of the predominant pattern was determined. To do so, the mean percentages per pattern were calculated for every case from the evaluation data of the five pulmonary pathologists. Subsequently, a consensus predominant pattern was assigned to each case on the basis of these data.

Statistical analysis

For the comparison of percentages of growth patterns, Pearson's correlation was used. The significance of differences in correlation coefficients between patterns was calculated by an ANOVA. For the comparison of κ-values between slide number sub-groups, the Mann–Whitney U-test was applied. All statistics and graphs were calculated with PASW 19 (IBM, Ehningen, Germany) and GraphPad Prism 4.03 (GraphPad Software, La Jolla, CA, USA).

RESULTS

Pattern frequency and slides

For the calculation of the number of tumour-containing slides only those slides that contained >10% tumour tissue were counted. Hence, a mean number of four tumour-bearing slides were available per case (range 1–12). The distribution of slide numbers bearing tumours per case is given in figure 2.

Correlation of pattern percentages

By calculating the mean overall frequency of patterns assigned by the five pulmonary pathologists, we found that the most frequent pattern in our cohort was acinar (32.9%), followed by solid (30.2%), lepidic (19.5%), papillary (10.4%) and micropapillary (7.1%) (fig. 2). When the correlation coefficients for the percentages per pattern were compared between pulmonary pathologists, we found considerable variations in dependability of the respective patterns (fig. 3). Correlation coefficients for pattern percentage between pulmonary pathologists were highest for the solid pattern (0.86) followed by the lepidic (0.78), acinar (0.61), micropapillary (0.60) and papillary (0.58) patterns. The differences in correlation coefficients between patterns were statistically significant ($p<0.001$). Correlation coefficients of the pathologist-in-training pattern percentages with the consensus frequency were considerably lower. Again, the correlation coefficient was highest for the solid pattern (0.64), followed by lepidic (0.50), acinar (0.46), papillary (0.42) and micropapillary (0.29). After training, correlation coefficients for the residents' pattern percentages with the consensus improved to values comparable to those

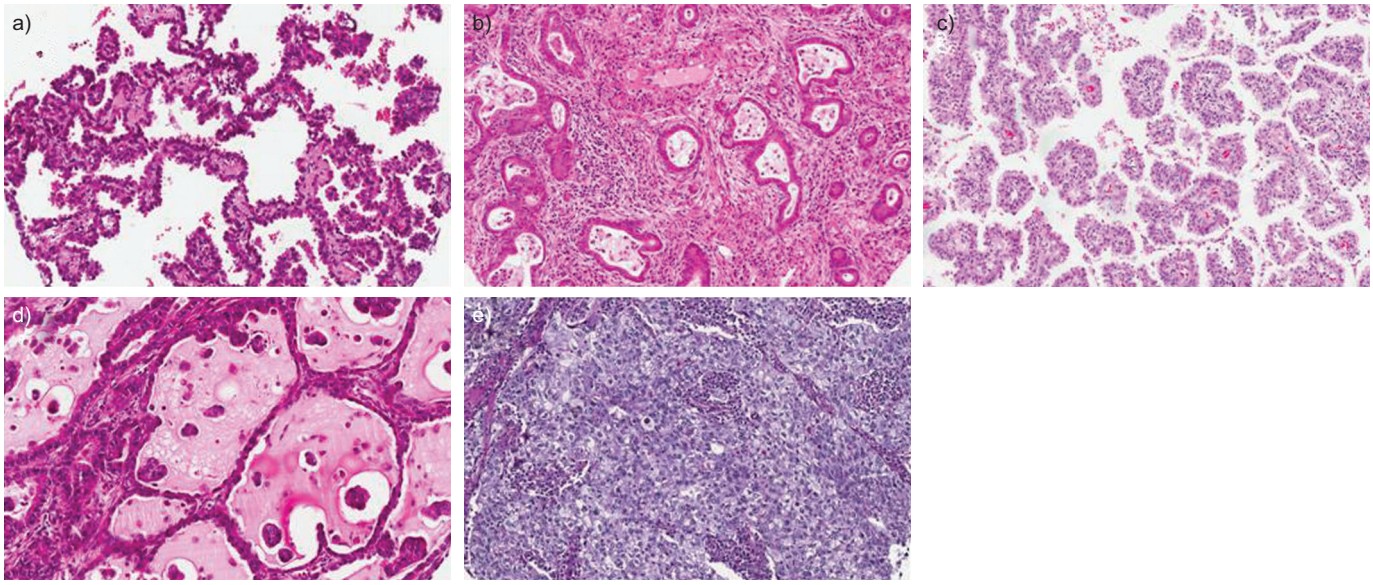

FIGURE 1. Exemplary photomicrographs of pulmonary adenocarcinoma (ADC) growth patterns. a–d) Haematoxylin and eosin and e) periodic acid–Schiff-stained pulmonary ADC with a) lepidic, b) acinar, c) papillary, d) micropapillary and e) solid growth patterns depicted. Original magnification 40×.

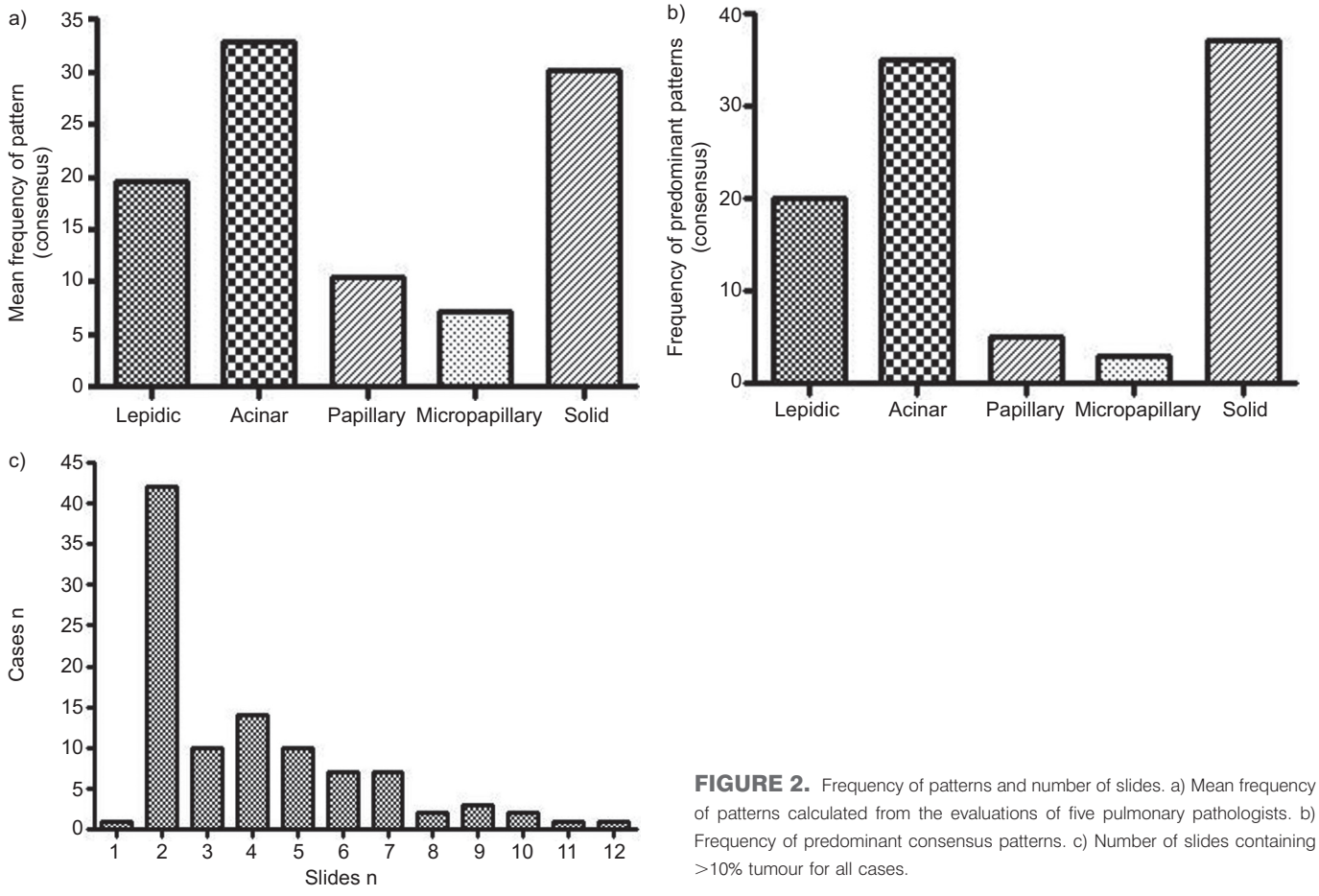

FIGURE 2. Frequency of patterns and number of slides. a) Mean frequency of patterns calculated from the evaluations of five pulmonary pathologists. b) Frequency of predominant consensus patterns. c) Number of slides containing >10% tumour for all cases.

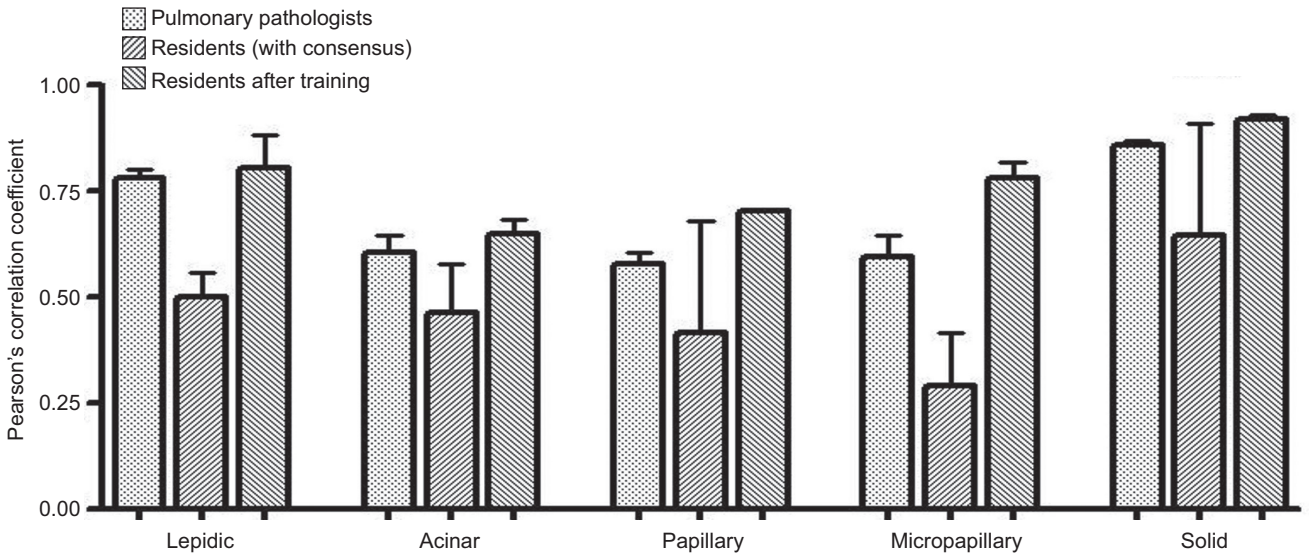

FIGURE 3. Pearson's correlation coefficients for pattern evaluation. Mean ± SD of Pearson's correlation coefficient is depicted separately for all patterns and evaluator groups.

of the pulmonary pathologists. Again, agreement for the solid pattern was highest (0.92), followed by lepidic (0.80), micropapillary (0.78), papillary (0.70) and acinar (0.65).

Comparison of the most prominent pattern

The most frequent predominant consensus pattern was solid (37%), followed by acinar (35%), lepidic (20%), papillary (5%) and micropapillary (3%) (fig. 2). κ-values for the predominant pattern for pulmonary pathologists from one institution (three comparisons) ranged between 0.51 and 0.72, and indicated moderate-to-substantial agreement. When predominant patterns were compared for all pulmonary pathologists, this resulted in κ-values between 0.44 and 0.72. For the comparison of the predominant pattern as evaluated by the untrained pathologists with the consensus of the pulmonary pathologists, κ-values were somewhat lower, and ranged between 0.38 and 0.47. However, after a training session, κ-values were elevated to 0.51 and 0.66 respectively. Repeated evaluation by the same reviewers resulted in very high κ-values of 0.79 and 0.87 (fig. 4), respectively.

Dependence of predominant pattern evaluation from the number of slides per case

To investigate whether κ-values for the predominant pattern were dependent on the number of slides evaluated, we created sub-groups of our cohort with one group containing those cases with one to three slides (n=53) and another group containing

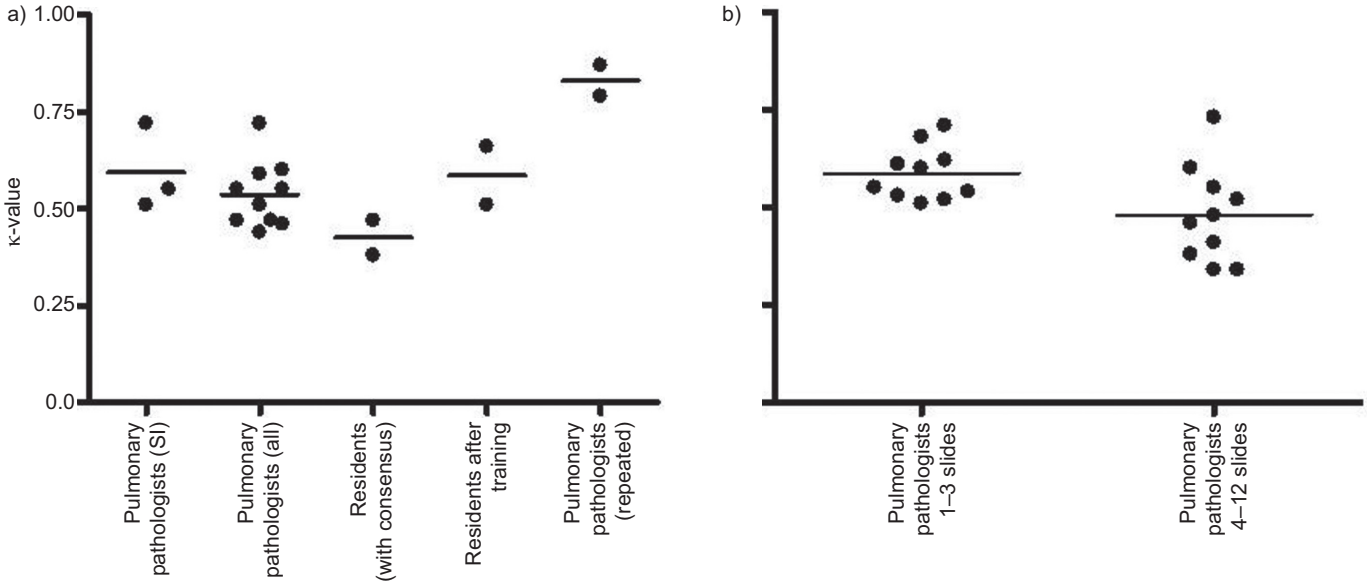

FIGURE 4. κ-values for the evaluation of the predominant pattern. a) κ-values are depicted separately for all evaluator groups; b) dependence of κ-values of pulmonary pathologists on the number of tumour-containing slides per case. SI: single institution.

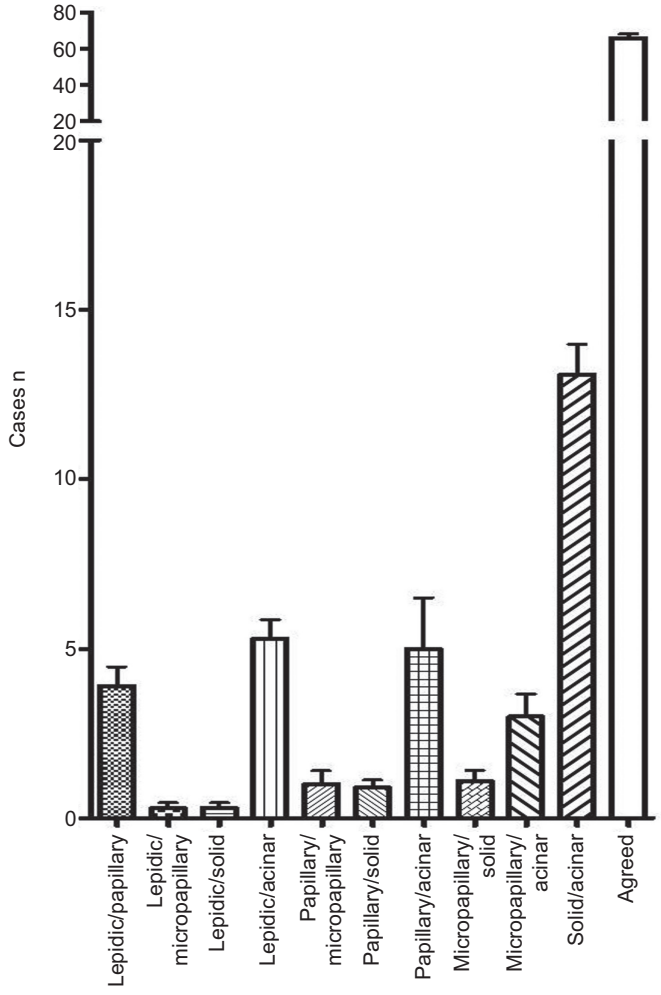

FIGURE 5. Mean number of cases with agreement/disagreement. Mean frequency of correct classifications and misclassifications for all possible pattern combinations from the evaluations of five pulmonary pathologists. Error bars represent SD.

the cases with four to 12 slides (n=47). Interestingly, κ-values for pulmonary pathologists were significantly worse (p=0.029) in the group with the high number of slides (mean 0.48) compared with the group with the low number of slides (mean 0.58) (fig. 4).

Discrepancies in the evaluation of selected patterns

To elucidate which were the most critical patterns to evaluate, we investigated how often disagreements in predominant patterns were observed between pulmonary pathologists for every possible combination. We found an agreement for the predominant pattern in a mean of 66 out of 100 cases. The most frequent absolute disagreement was between acinar and solid pattern with a mean of 13.1 cases, followed by the papillary/acinar (mean 5.3 cases) and lepidic/acinar (5.0 cases) combinations. Mean case numbers for lepidic/papillary, micropapillary/acinar and micropapillary/solid were 3.9, 3.0 and 1.1, respectively. Disagreement for the lepidic/micropapillary and lepidic/solid combinations was relatively rare (mean case numbers 0.3) (fig. 5).

Consequently, the absolute number of disagreed cases (10 comparisons between five pulmonary pathologists) was highest in those showing the acinar pattern (n=264), followed by solid (n=154), papillary (n=108), lepidic (n=98) and micropapillary (n=54) patterns (fig. 6).

Since the number of misclassifications is clearly biased by the overall frequencies in which the respective patterns can be observed, we then normalised the frequency of misclassifications per pattern. To do so, we calculated the number of disagreements for every pattern by summing up all disagreed cases in which the respective pattern was involved and divided this number by the number of cases classified as predominant for this pattern according to the consensus and by the number of possible comparisons (n=10). This allowed us to come up with a number of disagreed cases per predominant case per comparison for every pattern. This calculation showed that, in relative terms, indeed the most difficult patterns to evaluate were papillary (2.16) and micropapillary (1.8), followed by acinar (0.71), lepidic (0.49) and solid (0.44) (fig. 6).

DISCUSSION

The novel IASLC/ATS/ERS classification of pulmonary ADC [10] has been shown by us and others to have a strong and significant impact on overall, disease-specific and disease-free survival across all tumour stages [11, 13–15]. However, the question of the reproducibility of the identification of these patterns has not yet been addressed.

By performing a round-robin test under real-life conditions, we demonstrate that a reliable and reproducible pattern classification in lung ADC is possible. Agreement on the predominant growth pattern is moderate-to-good between pulmonary pathologists. Agreement on patterns was somewhat poorer for pathologists without pulmonary expertise in our study, but can be considerably improved by training; this argues in favour of training sessions for those pathologists not very accustomed to pulmonary pathology, in order to achieve better agreement.

While some groups have assessed the interobserver variability in the delineation of interstitial lung disease [16], diagnostic interobserver studies on lung neoplastic disorders are very scarce. One study focusing on the delineation of benign (scar lesion or reactive atypia) and malignant lesions of the lungs (variants of adenocarcinomas and squamous carcinomas including precursor lesions) by pulmonary pathologists reported κ-values of 0.65 and 0.81 (after training) [17]. For the delineation of small cell and large cell pulmonary carcinomas a κ-value of 0.4 has been reported [18]. To our knowledge, no other studies on the reproducibility of the IASLC/ATS/ERS classification have been published, which would potentially allow for data comparison. However, comparable studies have been published for interobserver studies on Gleason grading of prostatic adenocarcinomas by expert and general pathologists. In these studies, κ-values were strikingly similar to those found by us for pulmonary ADC [19–23].

Intraobserver variability in our study was quite low. However, since this has not been investigated before, this should be confirmed by other studies.

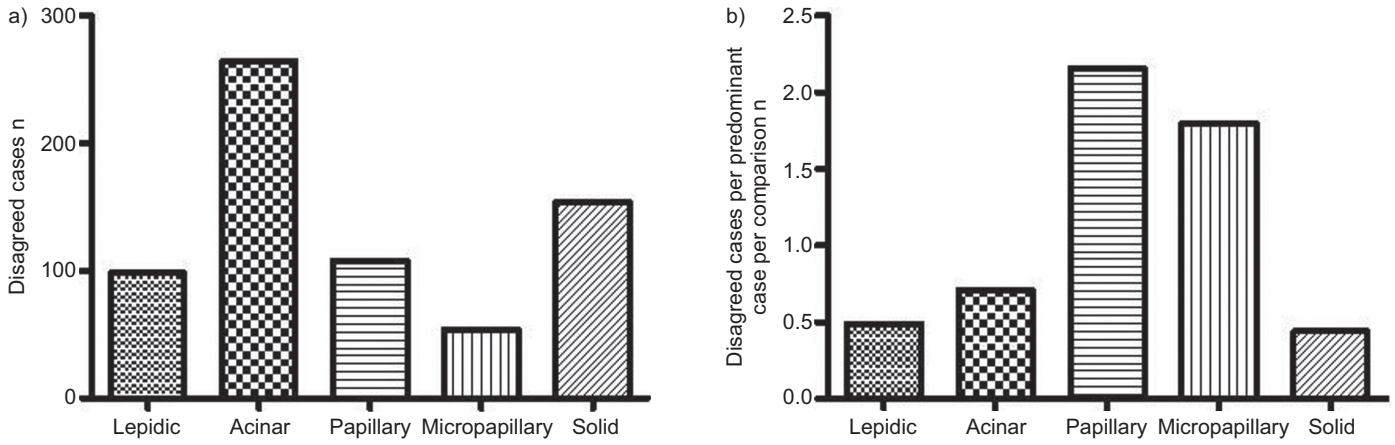

FIGURE 6. Misclassifications per pattern. a) Overall number of misclassifications involving the respective patterns; b) normalised number of misclassifications per predominant case and comparison for all patterns.

Unfortunately, in our dataset, κ-values decreased with increasing numbers of evaluated slides. Of course, increasing the number of analysed slides reveals more of the heterogeneity of pulmonary ADC and renders the evaluation of the predominant pattern by percentage more difficult. However, it is believed that accuracy of pattern classification increases when more tissue is investigated and, indeed, the number of patterns identified is dependent on the number of tumour-bearing slides available per case [13]. Because disagreement for the predominant pattern also increases with increasing slide numbers in some cases, it remains to be agreed upon which is the optimal number of slides to investigate and how to surgically dissect lung cancer specimens. However, we believe that the reproducibility of pattern classification in the high slide number sub-group will also increase as experience in this type of classification mounts and pathologists become more accustomed to this method.

The highest absolute numbers of cases with disagreement were those involving acinar, solid and lepidic patterns; however, these were also the predominant patterns with the highest frequency. Solid and lepidic patterns can be classified very reliably and the acinar pattern can be classified with moderate reliability. Nevertheless, the somewhat high overall rate of acinar/solid disagreements is alarming since prognosis and maybe even response to therapy might be considerably different in the two groups. We could recently provide first evidence that patients with predominant solid ADC might profit exceptionally from adjuvant radiotherapy. Furthermore, survival differences between stage I and stage II ADC and the risk of metastatic spread are significantly affected by the predominant pattern, which may also influence the choice of adjuvant chemotherapeutic treatment regimens in the future [13]. The same holds true for the relevant number of misclassified cases involving the lepidic pattern. Here again, the delineation is important since patients with lepidic predominant tumours have been shown to have a very good survival if compared to all other cases of ADC [11–14].

The absolute numbers of misclassified cases were medium or low for the papillary and micropapillary pattern, respectively. However, the normalised rate of misclassifications for these quite rare predominant growth patterns was highest for all patterns. If present in pulmonary ADC, papillary and micropapillary structures are often admixed in the same tumour area. While both patterns *per se* reached a fair interobserver agreement, a clear separation and definition of both patterns as the predominant one was one of the more complicated issues in this study. This underscores that this pattern combination is especially challenging to classify. These difficulties might also, at least partly, account for the striking differences in pattern frequencies for papillary and micropapillary predominant cases in the studies already published on this topic [11–14]. Taken together, the novel IASLC/ATS/ERS classification of pulmonary ADC is applicable with acceptable interobserver variability which is comparable to the interobserver variability of histomorphology-based classification schemes of other tumour entities already used in clinical decision making. A more precise definition of papillary and micropapillary criteria as well as more detailed criteria to separate acinar and solid pattern tumours may be helpful to reach an even better interobserver agreement. Additional efforts are needed to harmonise the application of this novel and clinically important classification scheme of pulmonary ADC.

STATEMENT OF INTEREST

A statement of interest for A. Stenzinger can be found at www.erj.ersjournals.com/site/misc/statements.xhtml

ACKNOWLEDGEMENTS

We gratefully acknowledge J. Schmitt (Institute for Pathology, Heidelberg University, Heidelberg, Germany) for excellent technical assistance.

REFERENCES

1 Jemal A, Siegel R, Xu J, *et al.* Cancer statistics, 2010. *CA Cancer J Clin* 2010; 60: 277–300.
2 Rami-Porta R, Bolejack V, Goldstraw P. The new tumor, node, and metastasis staging system. *Semin Respir Crit Care Med* 2011; 32: 44–51.
3 Kwak EL, Bang YJ, Camidge DR, *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med* 2010; 363: 1693–1703.
4 Rosell R, Moran T, Queralt C, *et al.* Screening for epidermal growth factor receptor mutations in lung cancer. *N Engl J Med* 2009; 361: 958–967.

5 Langer CJ, Besse B, Gualberto A, *et al.* The evolving role of histology in the management of advanced non-small-cell lung cancer. *J Clin Oncol* 2010; 28: 5311–5320.
6 Tomaszek SC, Huebner M, Wigle DA. Prospects for molecular staging of non-small-cell lung cancer from genomic alterations. *Expert Rev Respir Med* 2010; 4: 499–508.
7 Coindre JM. Grading of soft tissue sarcomas: review and update. *Arch Pathol Lab Med* 2006; 130: 1448–1453.
8 Cross SS. Grading and scoring in histopathology. *Histopathology* 1998; 33: 99–106.
9 Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J Urol* 1974; 111: 58–64.
10 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.
11 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.
12 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: 1155–1162.
13 Warth A, Muley T, Meister M, *et al.* The novel histologic IASLC/ATS/ERS classification of invasive pulmonary adenocarcinoma is a stage-independent predictor of survival. *J Clin Oncol* 2012; 30: 1438–1446.
14 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: 653–664.
15 Russel J, Stein A, Behrmann C, *et al.* Inflammatory lesions of the peritoneum mimic carcinomatosis after treatment with intravenous chemotherapy and intraperitoneal catumaxomab. *J Clin Oncol* 2011; 29: e644–646.
16 Thomeer M, Demedts M, Behr J, *et al.* Multidisciplinary interobserver agreement in the diagnosis of idiopathic pulmonary fibrosis. *Eur Respir J* 2008; 31: 585–591.
17 Thunnissen FB, Kerr KM, Brambilla E, *et al.* EU–USA pathology panel for uniform diagnosis in randomised controlled trials for HRCT screening in lung cancer. *Eur Respir J* 2006; 28: 1186–1189.
18 den Bakker MA, Willemsen S, Grunberg K, *et al.* Small cell carcinoma of the lung and large cell neuroendocrine carcinoma interobserver variability. *Histopathology* 2010; 56: 356–363.
19 Netto GJ, Eisenberger M, Epstein JI. Interobserver variability in histologic evaluation of radical prostatectomy between central and local pathologists: findings of TAX 3501 multinational clinical trial. *Urology* 2011; 77: 1155–1160.
20 Griffiths DF, Melia J, McWilliam LJ, *et al.* A study of Gleason score interpretation in different groups of UK pathologists; techniques for improving reproducibility. *Histopathology* 2006; 48: 655–662.
21 Melia J, Moseley R, Ball RY, *et al.* A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* 2006; 48: 644–654.
22 Glaessgen A, Hamberg H, Pihl CG, *et al.* Interobserver reproducibility of modified Gleason score in radical prostatectomy specimens. *Virchows Arch* 2004; 445: 17–21.
23 Burchardt M, Engers R, Muller M, *et al.* Interobserver reproducibility of Gleason grading: evaluation using prostate cancer tissue microarrays. *J Cancer Res Clin Oncol* 2008; 134: 1071–1078.