HHS Public Access

Author manuscript

Arch Pathol Lab Med. Author manuscript; available in PMC 2018 January 27.

Published in final edited form as:

Arch Pathol Lab Med. 2013 January ; 137(1): 32–40. doi:10.5858/arpa.2012-0033-OA.

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

Juneko E. Grilley-Olson, MD[*], **D. Neil Hayes, MD, MPH**[*,#], **Dominic T. Moore, MS, MPH**, **Kevin O. Leslie, MD**, **Matthew D. Wilkerson, PhD**, **Bahjat F. Qaqish, MD, PhD**, **Michele C. Hayward, RD**, **Christopher R. Cabanski, PhD**, **Xiaoying Yin, MD**, **Mark A. Socinski, MD**, **Thomas E. Stinchcombe, MD**, **Leigh B. Thorne, MD**, **Timothy Craig Allen, MD**, **Peter M. Banks, MD**, **Mary B. Beasley, MD**, **Alain C. Borczuk, MD**, **Philip T. Cagle, MD**, **Rebecca Christensen, MD**, **Thomas V. Colby, MD**, **Georgean G. Deblois, MD**, **Göran Elmberger, MD**, **Paolo Graziano, MD**, **Craig F. Hart, MD**, **Kirk D. Jones, MD**, **Diane M. Maia, MD**, **C. Ryan Miller, MD, PhD**, **Keith V. Nance, MD**, **William D. Travis, MD**, and **William K. Funkhouser, MD, PhD**[#]

Departments of Medicine, Division of Hematology-Oncology (Drs Grilley-Olson, Hayes, and Stinchcombe), The Lineberger Comprehensive Cancer Center (Drs Grilley-Olson, Hayes, Moore, Wilkerson, Qaqish, Hayward, Cabanski, Yin, Stinchcombe. Miller, and Funkhouser), and Pathology (Drs Thorne, Miller, and Funkhouser), University of North Carolina School of Medicine, Chapel Hill; the Department of Pathology, Mayo Clinic Arizona, Scottsdale (Drs Leslie and Colby); the Department of Statistics and Operations Research, University of North Carolina, Chapel Hill (Dr Cabanski); the Division of Hematology-Oncology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (Dr Socinski); the Department of Pathology, University of Texas Health Science Center, Tyler (Dr Allen); the Department of Pathology, Carolinas Medical Center, Charlotte, North Carolina (Dr Banks); the Department of Pathology, Mount Sinai Medical Center, New York, New York (Dr Beasley); the Department of Pathology, Columbia University Medical Center, New York (Dr Borczuk); the Department of Pathology, The Methodist Hospital System, Houston, Texas (Dr Cagle); the Department of Pathology, Norton Healthcare System, Louisville, Kentucky (Dr Christensen); the Department of Pathology, Commonwealth Laboratory Consultants, Richmond, Virginia (Dr Deblois); the Department of Pathology, Karolinska University Hospital, Stockholm, Sweden (Dr Elmberger); the Department of Pathology, C. Forlanini Hospital, Rome, Italy (Dr Graziano); the York Pathology Associates, Rock Hill, South Carolina (Dr Hart); the Department of Pathology, University of California, San Francisco (Dr Jones); the Kingsley Lane

Reprints: D Neil Hayes, MD, MPH, Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, 450 West Dr, CB 7295, Chapel Hill, NC 27599-7295 (hayes@med.unc.edu); William K Funkhouser, MD, PhD, Department of Pathology and Laboratory Medicine, University of North Carolina, CB 7525, Chapel Hill, NC 27599-7525 (WFunkhou@unch.unc.edu).

[*]These authors contributed equally to this manuscript.

[#]Co-corresponding authors.

The authors have no relevant financial interest in the products or companies described in this article.

Presented in part at the Metastatic Lung Session of the 45th Annual Meeting of the American Society of Clinical Oncology; May 29, 2009, to June 2, 2009; Orlando, Florida. Presented in part as a poster at the Pathology Session of the 13th World Conference on Lung Cancer; July 31, 2009, to August 4, 2009; San Francisco, California. Presented in part at the Pulmonary Pathology Society Meeting; June 24–26, 2009; Portland, Oregon.

Pathology Associates, Norfolk, Virginia (Dr Maia); the Department of Pathology, Rex Healthcare, Raleigh, North Carolina (Dr Nance); and the Department of Pathology, Memorial Sloan Kettering Cancer Center, New York (Dr Travis)

Abstract

Context—Precise subtype diagnosis of non–small cell lung carcinoma is increasingly relevant, based on the availability of subtype-specific therapies, such as bevacizumab and pemetrexed, and based on the subtype-specific prevalence of activating epidermal growth factor receptor mutations.

Objectives—To establish a baseline measure of inter-observer reproducibility for non–small cell lung carcinoma diagnoses with hematoxylin-eosin for the current 2004 World Health Organization classification, to estimate interobserver reproducibility for the therapeutically relevant squamous/nonsquamous subsets, and to examine characteristics that improve interobserver reproducibility.

Design—Primary, resected lung cancer specimens were converted to digital (virtual) slides. Based on a single hematoxylin-eosin virtual slide, pathologists were asked to assign a diagnosis using the 2004 World Health Organization classification. Kappa statistics were calculated for each pathologist-pair for each slide and were summarized by classification scheme, pulmonary pathology expertise, diagnostic confidence, and neoplastic grade.

Results—The 12 pulmonary pathology experts and the 12 community pathologists each independently diagnosed 48 to 96 single hematoxylin-eosin digital slides derived from 96 cases of non–small cell lung carcinoma resection. Overall agreement improved with simplification from the comprehensive 44 World Health Organization diagnoses ($\kappa = 0.25$) to their 10 major header subtypes ($\kappa = 0.48$) and improved again with simplification into the therapeutically relevant squamous/nonsquamous dichotomy ($\kappa = 0.55$). Multivariate analysis showed that higher diagnostic agreement was associated with better differentiation, better slide quality, higher diagnostic confidence, similar years of pathology experience, and pulmonary pathology expertise.

Conclusions—These data define the baseline diagnostic agreement for hematoxylin-eosin diagnosis of non–small cell lung carcinoma, allowing future studies to test for improved diagnostic agreement with reflex ancillary tests.

The diagnosis of non–small cell lung carcinoma (NSCLC) histologic subtype is the current gold standard for appropriate selection of chemotherapy, affecting the safety of bevacizumab[1] and the efficacy of pemetrexed.[2] The efficacy of epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors[3] is higher in patients with activating *EGFR* gene mutations,[4] present in 10% to 20% of lung adenocarcinoma (AD),[5] but few or no lung squamous carcinoma (SC).[6] Here, we estimate pathologists' diagnostic agreement by measuring interobserver reproducibility (IOR) for hematoxylin-eosin (H&E) diagnosis of NSCLC subtypes in resected specimens using the 2004 World Health Organization classification (2004-WHO).[7]

Four WHO lung cancer classifications have been published: 1967,[8] 1982,[9] 1999,[10] and 2004.[7] These classifications are based primarily on light microscopic evaluation of H&E-stained sections from resected neoplasms. Incremental refinements between editions have included reclassification for some disease entities (eg, solid AD with mucin production),

recognition of new disease entities (eg, large cell neuroendocrine carcinoma), fine-tuning of diagnostic criteria, and correlation with clinical, radiographic, immunohistochemical, and molecular variables.

Diagnostic agreement can be estimated by measuring percentage agreement or by calculating a κ statistic, which accounts for chance agreement. The κ statistic ranges from complete disagreement (κ = −1.0) to complete agreement (κ = 1.0), with a target minimum for clinical testing of 0.7.[11] Although the WHO classification system is complex, studies typically simplify categories.[12–24] Four studies assessed H&E-only IOR for NSCLC. Using the 1967 WHO classification, Feinstein et al[17] found 95% to 98% agreement for epidermoid and AD when well differentiated (WD), but only 58% to 60% agreement when poorly differentiated (PD). With the 1982 WHO classification, Hanai et al[20] and Yamamoto et al[24] reported 77% to 100% and 97% to 98% agreement, respectively. Burnett et al[12,13] reported κ = 0.28 to 0.30 for SC and AD with modest improvement when mucin stains were provided. Other IOR studies[14,15,18,19,22] are not directly comparable to the current study because they mix H&E-only diagnoses with diagnoses using both H&E and mucin stains. Employing the 1999 WHO classification, Colby et al[16] found dominant cell–type agreement in 74% to 82% of NSCLC/small cell lung cancer cases, with an overall κ of 0.65 to 0.74. No published IOR studies were found with the 2004-WHO classifications.

We designed this baseline study to measure the IOR (agreement) for diagnosis of resected NSCLC. Using the current 2004-WHO, we evaluated the IOR for the H&E diagnosis of NSCLC by comparing 24 pathologists' diagnoses of representative, digital H&E slides from 96 resected lung cancers. We report IORs for the complete 2004-WHO classification of 44 diagnoses (44DC) and estimate IORs for the classification's 10 major diagnostic categories (10DC) and the clinically relevant squamous/nonsquamous (SC/non-SC) classes (Table 1). We also report the effect of pathologists' practice settings, expertise in lung pathology, years of experience, confidence in the H&E diagnosis, slide quality, and carcinoma grade on IOR. This study is the first, to our knowledge, to measure the agreement of NSCLC H&E diagnoses for the entire current 2004-WHO, to estimate IOR for the therapeutically relevant SC/non-SC classes, and to demonstrate the utility of digital slide review.

METHODS

Sample Selection and Study Population

Sequential, surgically resected, primary NSCLCs (n = 96) collected at the University of North Carolina (Chapel Hill) from 1997–2007 were identified. Single diagnostic blocks used in the original pathologic diagnosis were recut and stained with H&E and were scanned using an Aperio ScanScope slide scanner (Aperio Technologies, Vista, California) into virtual slides viewable at magnifications equivalent to ×2 to ×20 objectives (×40 magnifier). Snapshot jpeg images (×2 and ×20) were created from unselected, central areas of the virtual slides. Grades were based on the original pathologic diagnosis. Small cell lung cancer, metastases, and normal specimens were excluded.

Increasing the number of pathologists increases the generalizability of the conclusions, and increasing the number of reviewed slides decreases the standard error around the κ estimate

of IOR.[25] To balance these considerations, we recruited 12 expert lung pathologists from the Pulmonary Pathology Society and 12 community pathologists. Each pathologist reviewed two random sets of 24 slides of the total 96 slides. Some pathologists elected to review all 96 slides.

Survey Content

Using DVDs containing virtual slides and Internet-based snapshots, pathologists recorded their 2004-WHO diagnoses onto an Internet-based survey. Pathologists were free to base their diagnoses on the virtual slide and/or the snapshot images. For each slide, pathologists reported diagnosis, quality of slide image, diagnostic confidence, and any additional comments. Pathologists' personal identifiers were removed by a designated data manager, but linked demographic information was retained, including years in practice and surgical pathology fellowship (yes/no), as well as whether the participant was an expert lung pathologist or a community pathologist. The study was approved by the University of North Carolina Institutional Review Board.

Statistical Methods

The Cohen[26] simple κ statistic was used to measure agreement among the 222 pathologist-pairs, from combinations of 24 pathologists. Pathologists' 44DC were collapsed into 10 DC and then into SC/non-SC categories (Table 1).[27]

Bootstrap methods[28,29] (including block bootstrapping) were used to calculate standard errors (standard deviations of the bootstrapped means), through which, 95% confidence intervals (CI) for the (weighted) mean κ statistics were calculated. Subgroup κ statistics were calculated along with their (bootstrap) 95% CI.

Exploratory analyses were performed using logistic regression modeling to examine possible associations of pathologist, slide, and tumor factors on the probability of agreement. The dependent variable of agreement on a diagnosis for a particular slide by a pathologist-pair was scored as agreement or disagreement. A c-index[30] was used to measure and compare the levels of association for both univariable and multivariable models. The covariates that were evaluated relating to the pathologists included expertise, practice setting, and years of diagnostic experience (both the sum of their combined experience, and the absolute values of the difference in their years of experience). We distinguished between tumor factors (inherent to the entire case as diagnosed by the original pathologist) and slide factors (inherent to the image being considered). Tumor factor covariates included pathologic diagnosis and original neoplastic grade. Slide factor covariates included confidence in diagnosis and image quality. In our logistic regression analyses, we dichotomized diagnosis as SC versus non-SC, grade as WD versus moderately differentiated (MD) versus PD, and confidence as *high* or *not high*. Because of the exploratory nature of the analysis, we did not adjust for the dependencies among slides and pathologists. Odds ratios with 95% CIs are given for these covariates of interest (Table 4).

Analyses were performed using both SAS (Version 9.2; SAS Institute, Inc, Cary, North Carolina) and R statistical software (R Development Core Team 2008).[31]

RESULTS

Twelve of 13 expert lung pathologists (92%) and 12 of 13 community pathologists (92%) agreed to participate in the study. A surgical pathology fellowship had been completed by 16 of 24 pathologists (67%). A median of 17 years (range, 1–36 years) of posttraining experience was reported (Table 2). Based on the 24 study pathologists reviewing random allocations of 48 to 96 slides, a comprehensive 1:1 matching of pathologists' pairwise agreements resulted in a total of 222 unique "pathologist-pairs" and 7130 unique slide viewings ("slide-pairs") reviewed by the pathologist-pairs. Slide-pairs (2 pathologists' diagnoses of a single slide) formed the fundamental unit by which we measured agreement.

All virtual slides contained cancer. All (96 of 96; 100%) low-power and 94% (90 of 96) of high-power jpeg snapshot images contained representative fields of the same neoplasm. Six percent (6 of 96) of the high-power jpeg snapshot images did not contain representative fields of the neoplasm seen in the low-power jpeg snapshot image. The IORs for pathologists who used primarily jpegs or both are similar with or without elimination of the 6 cases with nonrepresentative high-power jpeg images.

Four out of 24 pathologists (17%) experienced technical challenges in use of the large DVD virtual slide files and retrospectively reported using jpegs exclusively or a mixture of jpegs and DVDs. The IORs for pathologists who primarily used DVDs were similar to those who used primarily jpegs or both versions (data not shown).

On average, pathologists rated 91% of the diagnostic images of sufficient quality for diagnosis, with little agreement on which slides were of low quality. Quality was uniformly scored as acceptable in 37 of 96 (39%), with an additional 32 slides (33%) receiving only one unacceptable quality rating. Pathologists assigned confidence in their diagnoses as follows: high, 52%; moderate, 40%; and poor, 8% (Table 2).

The distribution of the original and study diagnoses were AD, 35% and 36%; SC, 35% and 31%; adenosquamous, 13% and 3%; large cell, 9% and 17%; miscellaneous, 6% and 7%; sarcomatoid carcinoma, 1% and 4%; and carcinoid, 1% and 2%, respectively. Based on the original pathologic grade, slides were 3% WD, 54% MD, and 43% PD (Table 2).

Overall, the IOR for H&E diagnoses for the entire 2004-WHO classification system (44DC), was $\kappa = 0.25$ (95% CI, 0.23–0.26) (Figure 1; Table 3). The 44DC κ statistics improved with simplification into 10DC (overall $\kappa = 0.48$), and again into the SC/non-SC classes (overall $\kappa = 0.55$; 95% CI, 0.53–0.58) and into the AD/non-AD classes (overall $\kappa = 0.59$; 95% CI, 0.57–0.61). Table 3 shows the variability of IOR as a function of diagnostic confidence, pulmonary pathology expertise, and neoplastic grade. The IOR varied most widely as a function of the pathologist's confidence in his or her H&E diagnosis. For each classification and level of expertise, IOR was higher when diagnostic confidence was higher. Overall, IOR improved by simplifying 44DC (high confidence $\kappa = 0.38$, moderate confidence $\kappa = 0.15$) into 10DC (high confidence $\kappa = 0.69$, moderate confidence $\kappa = 0.31$) and again into SC/non-SC classes (high confidence $\kappa = 0.78$, moderate confidence $\kappa = 0.28$).

For each classification (44DC, 10DC, dichotomous), IOR was higher when pulmonary pathology expertise was higher (Table 3). The IOR improved by simplifying the classification from 44DC (expert $\kappa = 0.30$, community $\kappa = 0.19$) into 10DC (expert $\kappa = 0.55$, community $\kappa = 0.36$), and again into SC/non-SC classes (expert $\kappa = 0.64$, community $\kappa = 0.41$) and AD/non-AD classes (expert $\kappa = 0.69$, community $\kappa = 0.46$).

For each classification (44DC, 10DC, dichotomous), IOR was higher when carcinomas were better differentiated (Table 3). The IOR improved by simplifying the 44DC (WD/MD $\kappa = 0.27$; PD $\kappa = 0.22$) into 10DC (WD/MD $\kappa = 0.52$; PD $\kappa = 0.41$) and again into the SC/non-SC (WD/MD $\kappa = 0.60$; PD $\kappa = 0.46$) and AD/non-AD (WD/MD $\kappa = 0.64$; PD $\kappa = 0.48$) classes. When considering only the 3 WD slides (all non-SC), pathologists were in 100% diagnostic agreement.

Mean agreement of each study pathologist's diagnosis with the original pathologist's diagnosis ($\kappa = 0.52$) was comparable to the overall IOR for 10DC of $\kappa = 0.48$. To assess the effect of potential outliers, study pathologist-pairs were stratified by pairwise agreement quartiles. The top quartile approached the goal of $\kappa = 0.70$ for good clinical agreement, whereas the bottom quartile had fair agreement. We identified both expert and community pathologists in all agreement quartiles (data not shown).

Tumor, slide, and pathologist variables were evaluated for univariable and multivariable effect on SC/non-SC IOR (Table 4). All univariable and all but one multivariable predictor (cumulative pathologist experience) were statistically significant. Predictors for higher IOR included better-differentiated carcinomas, better slide quality, and higher diagnostic confidence. Pathologist diagnostic confidence was statistically associated with neoplastic grade, slide quality, experience, and expertise. Because confidence was highly associated with the perception of slide quality ($P < .001$), any effect of slide quality on interpretation is probably reflected in the data regarding diagnostic confidence.

Increasing difference in years of pathologist practice experience predicted decreased IOR. Roughly, a 10% decrease in agreement was found for every 5 years difference in practice experience. Increased cumulative pathologist practice experience predicted increased IOR, statistically significant by univariate analysis only, with a 3% increase in agreement for every 5 years of cumulative practice experience. Pulmonary pathology expertise in both pathologists of a pair predicted an increased IOR: expert pathologist-pairs had a 38% increase in the odds of agreement compared with community pathologist pairs. Pulmonary pathology expertise was highly correlated with confidence, such that the odds of agreement for expert pathologist-pairs showed a 21% increase after controlling for confidence, quality, and grade in multivariable analysis (Table 4). Figure 2 graphically summarizes many of the results. Some cases, particularly WD cases of SC and AD, were readily identified with high IOR by H&E alone.

COMMENT

Strengths of the Study

Non–small cell lung carcinoma subtyping has refined and improved survival of patients with advanced NSCLC.[2,32] We designed a comprehensive prospective study of H&E diagnostic agreement for NSCLC. Using the 2004-WHO, our data measure IOR for the entire 44DC and provide estimates for the parent 10DC and the therapeutically relevant SC/ non-SC classes (Table 1). These data evaluate factors that might predict IOR, including sums and differences in years of practice experience, expertise in lung pathology, slide quality, diagnostic confidence, and carcinoma grade.

We hypothesized that IOR for the H&E diagnosis of NSCLC subtypes according to the 2004-WHO would show a κ of 0.7, an agreed-upon, albeit arbitrary, target for minimal clinical test reproducibility. We found that overall IOR among study pathologists was fair (κ = 0.25) when using all 44DC, with improvement following collapse into the 10DC (κ = 0.48) or the therapeutically relevant SC/non-SC classes (κ = 0.55) (Table 3). The low κ for 44DC is not surprising because many of these diagnoses would not be made in practice without ancillary stains. Our 10DC IOR results appear similar in magnitude to studies of prior versions of the classification,[12,13,15,16,18–22,24] but direct comparison to historic studies is limited because the most methodologically similar study[12,13] used bronchial biopsies rather than resection specimens. Additionally, other studies used glass slides and simplified the classification system into major diagnostic categories rather than using the comprehensive diagnostic listings.

Our multivariate analysis shows that grade, slide quality, diagnostic confidence, difference in experience, and pulmonary pathology expertise are independent predictors of NSCLC H&E diagnostic agreement, although those methods do not account for the dependencies among the slide review. Controllable factors that may improve agreement include optimizing H&E slide quality and increasing lung pathology expertise.

Our data suggest an upper limit for IOR by H&E alone, mainly because of PD NSCLC lacking morphologic features of SC or AD.[19] Pathologist confidence in his or her H&E slide diagnosis, the most predictive factor for increased IOR, likely reflects a qualitative amalgamation of grade, slide quality, and expertise. Diagnostic agreement may improve with systematic definition and application of reflex stain panels for PD NSCLC. Providing histochemical (eg, mucin) and immunohistochemical (eg, thyroid transcription factor 1, p63, cytokeratin 5/6, and napsin A) phenotypes, as well as cytogenetic tests (echinoderm microtubule-associated proteinlike 4 [EML4]– anaplastic lymphoma kinase [ALK] translocation) and molecular tests (eg, *EGFR/KRAS/BRAF* mutations) to define molecular targets for therapy likely would have improved diagnostic agreement; this is an important question for follow-up studies.

The 2004-WHO continues to reward the lung cancer community with meaningful associations, such as *EGFR* mutations with AD,[33] and the *EML4-ALK* fusion oncogene with signet-ring histology.[34] The goal remains to incrementally improve diagnostic

classifications, criteria, and reflex ancillary tests to optimize agreement, as well as to report associated prognostic and predictive data to guide patient management.

Although detailed classification likely reflects underlying biology, κ statistics increase with a reduced number of classes; therefore, simplifying the morphologic classification should improve agreement. Pathology reports that include both the specific (44DC) diagnosis and parent (10DC) category may reduce confusion by treating clinicians regarding management of uncommon WHO diagnoses.

Potential Limitations

Although our data include 7130 slide-pairs drawn from an incident patient series of 96 cases, we recognize that the sample size was insufficient to represent all diagnostic entities in the 2004-WHO. Diagnoses were based on single H&E images, rather than complete cases (glass slides with ancillary stains), with a goal of establishing baseline κ statistics for the H&E diagnosis of NSCLC. Based on feedback from several pathologists at the time the study was designed, we determined that reviewing 48 to 96 entire cases would deter participation. Study pathologists' agreement with each other was similar to their agreement with the original pathology diagnosis, arguing that our study design reflects what would have been observed if the entire case had been reviewed. We intentionally provided only H&E sections, without pertinent clinical, radiographic, or ancillary stain data, other than the knowledge that the patient carried a diagnosis of NSCLC, to estimate IOR of 3 relevant NSCLC classifications (44DC, 10DC, SC/non-SC) under conditions in which each pathologist had exactly the same information: an H&E image only.

Several pathologists lacked familiarity with digital images or had concerns regarding image resolution, which may have compromised their diagnostic abilities. However, digital images control for any variation in the circulated images, a major advantage over the morphologic variation inevitable in 24 recut sections through a paraffin block. Although not readily employed in clinical practice, it is commonly used in teaching and research, including for The Cancer Genome Atlas.[35] Wider use of digital slides could facilitate timely accrual to trials requiring central pathology review and expedite expert review of challenging cases.

The IOR was similar among pathologists who primarily used DVDs versus jpegs or both (regardless of the 6 cases with nonrepresentative, high-power jpeg images). These data argue that IOR estimates were not affected by pathologist decision to use snapshots versus DVD images, or by the 6% of cases with nonrepresentative ×20 snapshots.

Our resected specimen results may be extrapolated to, but may not fully represent, small biopsies and fine-needle aspirates from patients with advanced NSCLC. Recently, the International Association for the Study of Lung Cancer/American Thoracic Society/ European Respiratory Society[36] published major changes in the lung AD subclassification, including guidelines for small-biopsy diagnosis, although those changes do not alter the distinction among the major 10DC subtypes, such as SC and AD.

The SC/non-SC categorization is not a feature of the WHO classification but, rather, was based on clinical and regulatory practice: pemetrexed has no proven efficacy in SC in any of

3 pivotal studies[2,37,38] contributing to the drug's approval in non-SC histology NSCLC, and bevacizumab is contraindicated in SC because of potential life-threatening hemorrhage.[1] Our study was executed in 2008, before the publication of pivotal studies related to pemetrexed and bevacizumab in journals not directed at pathologists. Nevertheless, we demonstrate that even a simple classification, such as SC/non-SC, is imperfect by H&E alone (SC/non-SC, experts, maximum $\kappa = 0.84$).

CONCLUSIONS

Management of advanced NSCLC is now critically dependent on precise histologic diagnoses. This study provides baseline estimates of the IOR for H&E diagnosis of NSCLC and shows that agreement is a function of pathologist experience, pulmonary pathology expertise, pathologist diagnostic confidence, slide quality, and carcinoma grade. Strict definition and application of diagnostic criteria may incrementally improve IOR for H&E diagnosis of NSCLC, but major improvements in NSCLC IOR will likely depend on systematic integration of validated histochemical, immunohistochemical, and molecular methods. We recommend reporting the major (10DC) diagnostic category along with the specific (44DC) WHO diagnosis, thereby providing the groundwork for further therapeutic advances while reducing the potential for clinical confusion in how to manage unusual NSCLC cases. Our findings define a baseline measure for NSCLC H&E diagnostic agreement, to which future studies determining incremental benefits of reflex ancillary tests at the protein, cytogenetic, and molecular levels may be compared.

Acknowledgments

Research was supported by a grant from the Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill.

References

1. Johnson DH, Fehrenbacher L, Novotny WF, et al. Randomized phase II trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non– small-cell lung cancer. J Clin Oncol. 2004; 22(11): 2184–2191. [PubMed: 15169807]
2. Scagliotti GV, Parikh P, von Pawel J, et al. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non–small-cell lung cancer. J Clin Oncol. 2008; 26(21):3543–3551. [PubMed: 18506025]
3. Hirsch FR, Spreafico A, Novello S, Wood MD, Simms L, Papotti M. The prognostic and predictive role of histology in advanced non–small cell lung cancer: a literature review. J Thorac Oncol. 2008; 3(12):1468–1481. [PubMed: 19057275]
4. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009; 361(10):947–957. [PubMed: 19692680]
5. Rosell R, Moran T, Queralt C, et al. Screening for epidermal growth factor receptor mutations in lung cancer. N Engl J Med. 2009; 361(10):958–967. [PubMed: 19692684]
6. Marchetti A, Martella C, Felicioni L, et al. EGFR mutations in non–small-cell lung cancer: analysis of a large series of cases and development of a rapid and sensitive method for diagnostic screening with potential implications on pharmacologic treatment. J Clin Oncol. 2005; 23(4):857–865. [PubMed: 15681531]

7. Travis, WD., Brambilla, E., Muller-Hermelink, HK., Harris, CC. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus, and Heart. 3rd. Vol. 10. Lyon, France: IARC Press; 2004. World Health Organization Classification of Tumours
8. Kreyberg, L., Liebow, AA., Uehlinger, EA. Histological Typing of Lung Tumours. 1st. Geneva, Switzerland: World Health Organization; 1967. International Histological Classification of Tumours
9. The World Health Organization histological typing of lung tumours: second edition. Am J Clin Pathol. 1982; 77(2):123–136. [PubMed: 7064914]
10. Travis, WD., Colby, TV., Corrin, B., Shimosato, Y., Brambilla, E., Sobin, LH. Histological Typing of Lung and Pleural Tumours. 3rd. Berlin, Germany: Springer-Verlag; 1999. World Health Organization International Histological Classification of Tumours
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]
12. Burnett RA, Howatson SR, Lang S, et al. Observer variability in histopathological reporting of non–small cell lung carcinoma on bronchial biopsy specimens. J Clin Pathol. 1996; 49(2):130–133. [PubMed: 8655678]
13. Burnett RA, Swanson Beck J, Howatson SR, et al. Observer variability in histopathological reporting of malignant bronchial biopsy specimens. J Clin Pathol. 1994; 47(8):711–713. [PubMed: 7962622]
14. Butler C, Samet JM, Humble CG, Sweeney ES. Histopathology of lung cancer in New Mexico, 1970–72 and 1980–81. J Natl Cancer Inst. 1987; 78(1):85–90. [PubMed: 3025506]
15. Campobasso O, Andrion A, Ribotta M, Ronco G. The value of the 1981 WHO histological classification in inter-observer reproducibility and changing pattern of lung cancer. Int J Cancer. 1993; 53(2):205–208. [PubMed: 8381110]
16. Colby TV, Tazelaar HD, Travis WD, Bergstralh EJ, Jett JR. Pathologic review of the Mayo Lung Project cancers [corrected]: is there a case for misdiagnosis or overdiagnosis of lung carcinoma in the screened group? Cancer. 2002; 95(11):2361–2365. [PubMed: 12436443]
17. Feinstein AR, Gelfman NA, Yesner R. Observer variability in the histopathologic diagnosis of lung cancer. Am Rev Respir Dis. 1970; 101(5):671–684. [PubMed: 4910640]
18. Field RW, Smith BJ, Platz CE, et al. Lung cancer histologic type in the surveillance, epidemiology, and end results registry versus independent review. J Natl Cancer Inst. 2004; 96(14):1105–1107. [PubMed: 15265973]
19. Ghandur-Mnaymneh L, Raub WA Jr, Sridhar KS, Albores-Saavedra J, Gould E, Duncan RC. The accuracy of the histological classification of lung carcinoma and its reproducibility: a study of 75 archival cases of adenosquamous carcinoma. Cancer Invest. 1993; 11(6):641–651. [PubMed: 8221196]
20. Hanai A, Whittaker JS, Tateishi R, Sobin LH, Benn RT, Muir CS. Concordance of histological classification of lung cancer with special reference to adenocarcinoma in Osaka, Japan, and the North-West Region of England. Int J Cancer. 1987; 39(1):6–9. [PubMed: 3793271]
21. Kreuzer M, Muller KM, Brachner A, et al. Histopathologic findings of lung carcinoma in German uranium miners. Cancer. 2000; 89(12):2613–2621. [PubMed: 11135223]
22. Sorensen JB, Hirsch FR, Gazdar A, Olsen JE. Interobserver variability in histopathologic subtyping and grading of pulmonary adenocarcinoma. Cancer. 1993; 71(10):2971–2976. [PubMed: 8387872]
23. Stang A, Pohlabeln H, Muller KM, Jahn I, Giersiepen K, Jockel KH. Diagnostic agreement in the histopathological evaluation of lung cancer tissue in a population-based case-control study. Lung Cancer. 2006; 52(1):29–36. [PubMed: 16476504]
24. Yamamoto S, Sobue T, Yamaguchi N, et al. Reproducibility of diagnosis and its influence on the distribution of lung cancer by histologic type in Osaka, Japan. Jpn J Cancer Res. 2000; 91(1):1–8. [PubMed: 10744038]
25. Altman, DG. Practical Statistics for Medical Research. Boca Raton, FL: Chapman & Hall/CRC; 1991.
26. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20:37–46.
27. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971; 76(5):378–382.

28. Liu RY, Singh K. Using i.i.d. bootstrap inference for general non-i.i.d. models. J Stat Plan Infer. 1995; 43(1–2):67–75.

29. DasGupta, A. SpringerLink: Asymptotic Theory of Statistics and Probability. New York, NY: Springer; 2008. Springer Texts in Statistics

30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. [PubMed: 7063747]

31. R Development Core. Team R: A Language and Environment for Statistical Computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; 2008.

32. Sandler A, Gray R, Perry MC, et al. Paclitaxel-carboplatin alone or with bevacizumab for non–small-cell lung cancer. N Engl J Med. 2006; 355(24):2542–2550. [PubMed: 17167137]

33. Sarkaria IS, Zakowski MF, Pham D, et al. Epidermal growth factor receptor signaling in adenocarcinomas with bronchioloalveolar components. Ann Thorac Surg. 2008; 85(1):216–223. [PubMed: 18154814]

34. Shaw AT, Yeap BY, Mino-Kenudson M, et al. Clinical features and outcome of patients with non–small-cell lung cancer who harbor EML4-ALK. J Clin Oncol. 2009; 27(26):4247–4253. [PubMed: 19667264]

35. The Cancer Genome Atlas Research Network. et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061–1068. [PubMed: 18772890]

36. 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. [PubMed: 21252716]

37. Ciuleanu T, Brodowicz T, Zielinski C, et al. Maintenance pemetrexed plus best supportive care versus placebo plus best supportive care for non–small-cell lung cancer: a randomised, double-blind, phase 3 study. Lancet. 2009; 374(9699):1432–1440. [PubMed: 19767093]

38. Scagliotti G, Hanna N, Fossella F, et al. The differential efficacy of pemetrexed according to NSCLC histology: a review of two phase III studies. Oncologist. 2009; 14(3):253–263. [PubMed: 19221167]

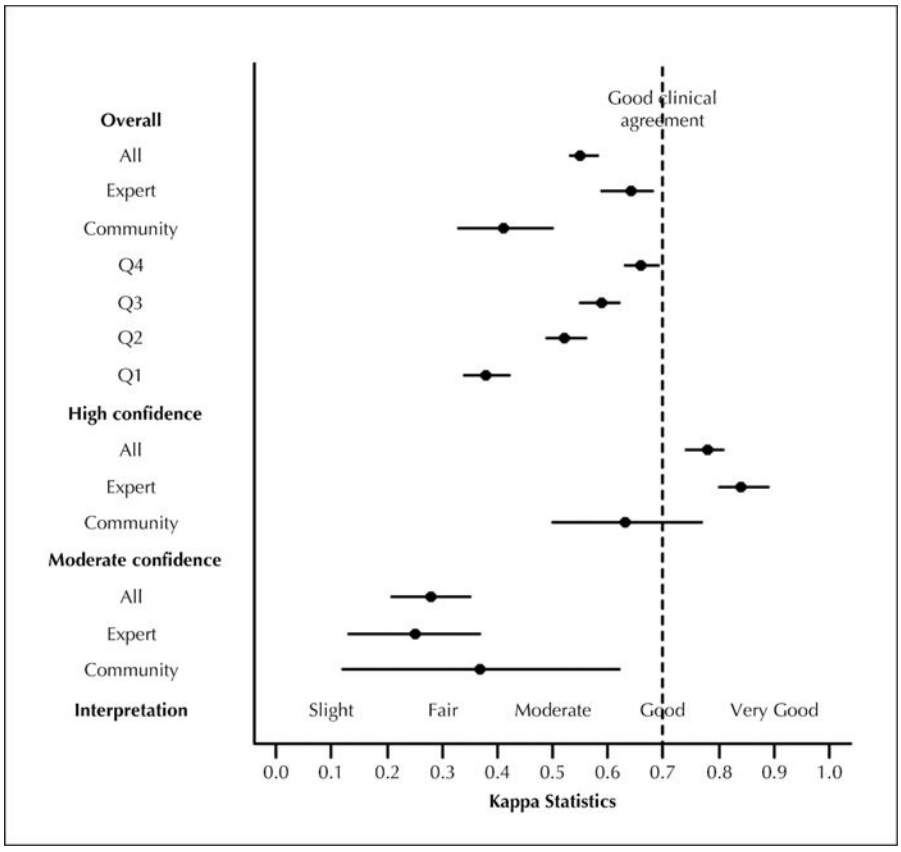

Figure 1.

Pathologist agreement for hematoxylin-eosin diagnosis of squamous versus non-squamous carcinoma. Abbreviations: All, all participant pathologists; Q1, first quartile; Q2, second quartile; Q3, third quartile; Q4, fourth quartile.

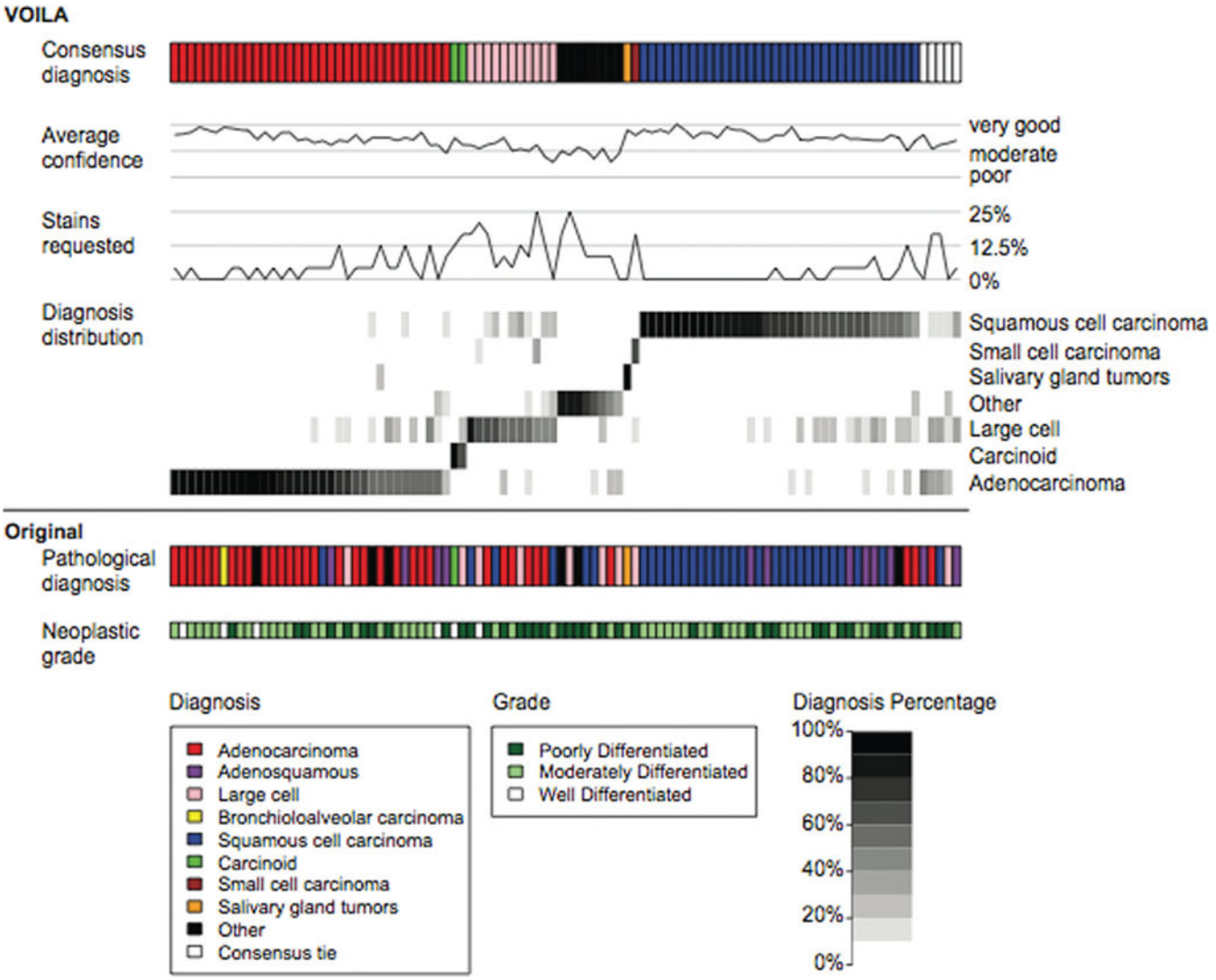

Figure 2.

Each column of the figure corresponds to 1 of the 96 hematoxylin-eosin (H&E) slides. The top row shows study pathologist consensus (the majority) diagnosis. Average confidence and request for immunohistochemical stains across all reviewers are plotted as a function of the slides. Diagnosis distribution represents a heat map of the fraction of times any given 10 diagnostic-class (10DC) diagnosis was made for each of the 96 slides. The bottom 2 rows show the original pathologic diagnosis and the original neoplastic grade.

Table 1

Adaptation of the World Health Organization (4th ed) Classification System[a] and Corresponding Simplified 10- and 2-Class Systems

44 Diagnostic Categories	10 Diagnostic Categories	Therapeutically Relevant 2 Classes: SC/Non-SC
MALIGNANT EPITHELIAL TUMORS		
Squamous cell carcinoma	Squamous carcinoma	Squamous carcinoma
Papillary		
Clear cell		
Small cell		
Basaloid		
Small cell carcinoma	Small cell carcinoma	Nonsquamous carcinoma (other NSCLC)
Combined small cell carcinoma		
Adenocarcinoma	Adenocarcinoma	
Adenocarcinoma, mixed subtype		
Acinar adenocarcinoma		
Papillary adenocarcinoma		
Bronchioloalveolar carcinoma		
Nonmucinous		
Mucinous		
Mixed nonmucinous and mucinous or indeterminate		
Solid adenocarcinoma with mucin production		
Fetal adenocarcinoma		
Mucinous (“colloid”) carcinoma/mucinous cystadenocarcinoma		
Signet-ring adenocarcinoma		
Clear cell adenocarcinoma		
Large cell carcinoma	Large cell carcinoma	
Large cell neuroendocrine carcinoma		
Combined large cell neuroendocrine carcinoma		
Basaloid carcinoma		
Lymphoepithelioma-like carcinoma		
Clear cell carcinoma		
Large cell carcinoma with rhabdoid phenotype		
Adenosquamous carcinoma	Adenosquamous carcinoma	
Sarcomatoid carcinoma	Sarcomatoid carcinoma	
Pleomorphic carcinoma		
Spindle cell carcinoma		
Giant cell carcinoma		
Carcinosarcoma		
Pulmonary blastoma		
Carcinoid tumor	Carcinoid tumor	Other carcinomas
Typical carcinoid		

44 Diagnostic Categories	10 Diagnostic Categories	Therapeutically Relevant 2 Classes: SC/Non-SC
Atypical carcinoid		
Salivary gland tumors	Salivary gland tumor	
Mucoepidermoid carcinoma		
Adenoid cystic carcinoma		
Epithelial-myoepithelial carcinoma		
MISCELLANEOUS TUMORS[b] (including mesenchymal tumors and lymphoproliferative tumors)	Miscellaneous tumors[b]	
METASTATIC TUMORS	Metastatic tumors	

Abbreviations: SC/non-SC, squamous versus nonsquamous carcinoma; NSCLC, non–small cell lung cancer.

[a] Data derived from Travis et al,[7] 2004.

[b] Modified: Mesenchymal tumors and lymphoproliferative tumors were added to miscellaneous category. Omitted: preinvasive lesions, benign epithelial tumors.

Table 2

Study Participant, Methodology, and Diagnosis Demographics and Statistics

Characteristics of the Study Participants	
Characteristics of the Pathologists	**Demographics, No. (%)**
Total number of participants/total approached	24/26 (92)
Sex, M	18 (75)
Surgical pathology fellowship, yes	16 (67)
Expert lung pathologist, yes	12 (50)
Years of experience:	
Minimum	1
Median	17
Maximum	36
Quality Assessment of Study Methodology	
Characteristics of the H&E Slides	**Frequency, %**
Slide of sufficient quality, yes	91
Confidence in assigned diagnosis:	
High	52
Moderate	40
Low	8
Request for special stains, yes	10

Distribution of Morphologic Diagnoses		
Morphology	**Original Diagnosis, %**	**Study Diagnosis, %**
Squamous cell	35%	36%
Adenocarcinoma	35%	31%
Adenosquamous	13	3
Large cell	9	17
Miscellaneous[a]	6	7
Sarcomatoid carcinoma	1	4
Carcinoid	1	2

Original Pathologic Grade Assigned to Study H&E Slides	
Original Pathologic Grade	**Results, %**
Not poorly differentiated:	57
Well differentiated	3
Moderately differentiated	54
Poorly differentiated	43

Abbreviation: H&E, hematoxylin-eosin.

[a]Category includes adenoid cystic, mucoepidermoid, spindle cell, basaloid non–small cell lung carcinoma, and malignant mesothelioma.

Table 3

κ Statistics by Pathologist and Diagnostic Category, Based on Hematoxylin-Eosin Diagnosis

Diagnostic Category	Reviewing Pathologist	Overall	High Confidence	Moderate Confidence	Well- and Moderately Differentiated	Poorly Differentiated
44DC	All	0.25	0.38	0.15	0.27	0.22
	Expert	0.30	0.41	0.15	0.31	0.28
	Community	0.19	0.37	0.19	0.24	0.13
10DC	All	0.48	0.69	0.31	0.52	0.41
	Expert	0.55	0.77	0.32	0.59	0.48
	Community	0.36	0.51	0.28	0.41	0.28
SC/non-SC	All	0.55	0.78	0.28	0.60	0.46
	Expert	0.64	0.84	0.25	0.68	0.53
	Community	0.41	0.63	0.37	0.46	0.32
AD/non-AD	All	0.59	0.74	0.40	0.64	0.48
	Expert	0.69	0.83	0.45	0.70	0.60
	Community	0.46	0.49	0.22	0.55	0.30

Abbreviations: 10DC, World Health Organization diagnostic classification system (Travis et al,[7] 2004) collapsed into 10 primary categories; 44DC, complete World Health Organization diagnostic classification system (Travis et al,[7] 2004); AD/non-AD, adenocarcinoma versus nonadenocarcinoma; all, all participant pathologists; community, community pathologists; expert, pulmonary pathologist experts; SC/non-SC, squamous versus nonsquamous carcinoma.

Table 4

Univariable and Multivariable Predictors of Interobserver Reproducibility in the Hematoxylin-Eosin (H&E) Diagnosis of Squamous Versus Nonsquamous Carcinoma

Variable	Univariable Predictors, OR (95% CI)	Multivariable Predictors, OR (95% CI)	Interpretation
Tumor factors			
Grade[a]	0.722 (0.646–0.808)	0.812 (0.714–0.922)	Decreased agreement in more-poorly differentiated tumors
H&E slide factors			
Quality[b]	0.573 (0.494–0.665)	0.689 (0.576–0.823)	Decreased agreement if one or both pathologists felt the slide was of low quality
Confidence[c]	2.21 (1.86–2.63)	2.02 (1.68–2.41)	Increased agreement if both pathologists were highly confident
Pathologist factors			
Difference in pathologist experience[d]	0.904 (0.864–0.945)	0.901 (0.860–0.944)	For every 5 y in experience difference, there is a 10% decrease in the odds of agreement
Cumulative pathologist experience[e]	1.03 (1.01–1.07)	1.03 (1.00–1.06)	For every 5 y of combined experience, there is a 3% increase in the odds of agreement
Pulmonary pathology expert[f]	1.38 (1.21–1.58)	1.21 (1.04–1.41)	Increased agreement if both are expert

Abbreviations: CI, confidence interval; H&E, hematoxylin-eosin; OR, odds ratio.

[a]Grade based on the original pathologic diagnosis (well-differentiated, < moderately differentiated, < poorly differentiated); the OR for 1 unit increase in grade.

[b]The OR comparing cases where one or more pathologists felt the quality of slide was insufficient to cases where both felt the quality was sufficient.

[c]The OR comparing cases where both pathologists had high confidence to any other confidence pairing.

[d]The OR for difference in experience between the pathologists per 5-year block.

[e]The OR for combined sum of experience for the pathologist pair per 5-year block.

[f]The OR comparing cases where both pathologists were experts compared with neither being expert.