

Interpathologist Diagnostic Agreement for Non–Small Cell Lung Carcinomas Using Current and Recent Classifications

William K. Funkhouser Jr, MD, PhD; D. Neil Hayes, MD, MPH; Dominic T. Moore, MPH, MS; W. Keith Funkhouser III, MS; Jason P. Fine, PhD; HeeJoon Jo, BA; Nana Nikolaishvili-Feinberg, PhD; Mervi Eeva, BS; Juneko E. Grilley-Olson, MD; Peter M. Banks, MD; Paolo Graziano, MD; Elizabeth L. Boswell, MD; Goran Elmberger, MD; Kirtee Raparia, MD; Craig F. Hart, MD; Lynette M. Sholl, MD; Norris J. Nolan, MD; Karen J. Fritchie, MD; Ersie Pouagare, MD; Timothy C. Allen, MD, JD; Keith E. Volmar, MD; Paul W. Biddinger, MD; Daniel T. Kleven, MD; Michael J. Papez, MD; Deborah V. Spencer, MD; Natasha Rekhtman, MD, PhD; Mari Mino-Kenudson, MD; Lida Hariri, MD, PhD; Brandon Driver, MD; Philip T. Cagle, MD

Accepted for publication March 26, 2018.

Published online April 30, 2018.

Supplemental digital content is available at www.archivesofpathology.org in the December 2018 table of contents.

From the Department of Pathology & Laboratory Medicine, University of North Carolina School of Medicine, Chapel Hill (Drs Funkhouser Jr and Banks); Lineberger Comprehensive Cancer Center, UNC, Chapel Hill, North Carolina (Drs Funkhouser Jr, Hayes, Nikolaishvili-Feinberg, and Grilley-Olson; Messrs Moore and Jo; and Ms Eeva); the Department of Medicine, UNC School of Medicine, Chapel Hill, North Carolina (Drs Hayes and Grilley-Olson); the Department of Computer Sciences, University of Wisconsin, Madison (Mr Funkhouser III); the Department of Biostatistics, UNC School of Public Health, Chapel Hill, North Carolina (Dr Fine); Medical Affairs, Ventana Medical Systems, Tucson, Arizona (Dr Banks); Unit of Pathology, Scientific Institute for Research and Health Care, San Giovanni Rotondo, Italy (Dr Graziano); the Department of Pathology, VA Medical Center, Durham, North Carolina (Dr Boswell); the Department of Medical Biosciences, Pathology, Umeå University Hospital, Umeå, Sweden (Dr Elmberger); the Department of Pathology, Kaiser-Permanente Hospital, Santa Clara, California (Dr Raparia); the Department of Pathology, Piedmont Medical Center, Rock Hill, South Carolina (Dr Hart); the Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts (Dr Sholl); the Department of Pathology, Suburban Hospital, Bethesda, Maryland (Dr Nolan); the Department of Pathology, Mayo Clinic, Rochester, Minnesota (Dr Fritchie); the Department of Pathology, VA Medical Center, Dayton, Ohio (Dr Pouagare); the Department of Pathology, University of Texas Medical Branch, Galveston (Dr Allen); the Department of Pathology, Rex Hospital, Raleigh, North Carolina (Dr Volmar); the Department of Pathology, Medical College of Georgia, Augusta (Drs Biddinger and Kleven); the Department of Pathology, Flagstaff Medical Center, Flagstaff, Arizona (Dr Papez); the Department of Pathology, VA Medical Center, Charleston, South Carolina (Dr Spencer); the Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York (Dr Rekhtman); the Department of Pathology, Massachusetts General Hospital, Boston (Drs Mino-Kenudson and Hariri); and the Department of Pathology & Genomic Medicine, Houston Methodist Hospital, Houston, Texas (Drs Driver and Cagle). Dr Allen is currently located at the Department of Pathology at University of Mississippi Medical Center, Jackson.

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

Corresponding author: William K. Funkhouser Jr, MD, PhD, Department of Pathology and Lab Medicine, CB No. 7525, UNC School of Medicine, Chapel Hill, NC 27599-7525 (email: Bill_Funkhouser@med.unc.edu).

• **Context.**—Measurement of interpathologist diagnostic agreement (IPDA) should allow pathologists to improve current diagnostic criteria and disease classifications.

Objectives.—To determine how IPDA for pathologists' diagnoses of non–small cell lung carcinoma (NSCLC) is affected by the addition of a set of mucin and immunohistochemical (IHC) stains to hematoxylin-eosin (H&E) alone, by recent NSCLC reclassifications, by simplification of these classifications, and by pathologists' practice location, pulmonary pathology expertise, practice duration, and lung carcinoma case exposure.

Design.—We used a Web-based survey to present core images of 54 NSCLC cases to 22 practicing pathologists for diagnosis, initially as H&E only, then as H&E plus mucin and 4 IHC stains. Each case was diagnosed according to published 2004, 2011, and 2015 NSCLC classifications. Cohen's kappa was calculated for the 231 pathologist pairs as a measure of IPDA.

Results.—Twenty-two pathologists diagnosed 54 NSCLC cases by using 4 published classifications. IPDA is significantly higher for H&E/mucin/IHC diagnoses than for H&E-only diagnoses. IPDA for H&E/mucin/IHC diagnoses is highest with the 2015 classification. IPDA is estimated higher after collapse of stated diagnoses into subhead or dichotomized classes. IPDA for H&E/mucin/IHC diagnoses with the 2015 World Health Organization classification is similar for community and academic pathologists, and is higher when pathologists have pulmonary pathology expertise, have more than 6 years of practice experience, or diagnose more than 100 new lung carcinoma cases per year.

Conclusions.—Higher IPDA is associated with use of mucin and IHC stains, with the 2015 NSCLC classification, and with pathologists' pulmonary pathology expertise, practice duration, and frequency of lung carcinoma cases.

(*Arch Pathol Lab Med.* 2018;142:1537–1548; doi: 10.5858/arpa.2017-0481-OA)

It is incumbent upon physicians of all specialties to continuously measure and improve the overall quality of patient care.¹ In addition to creating new knowledge

relevant to prevention, prognosis, and prediction of response to treatment of human diseases, pathologists should continuously improve diagnostic criteria and disease classifications to improve diagnostic test performance² by reducing diagnostic variability and thus optimizing inter-pathologist diagnostic agreement (IPDA). IPDA can be estimated by using accepted statistical measures of interrater agreement for nominal data that allow correction for random agreement, for example, the Cohen kappa, k .^{3–8} Kappa estimates the degree to which raters (in this case, pathologists) agree more often than predicted by chance alone. Kappa ranges from (+)1.0 (perfect interrater agreement) to 0 (random agreement) to (–)1.0 (perfect interrater disagreement). Kappa has been used by pathologists to test their abilities to distinguish benign from malignant precursor lesions,^{9,10} to distinguish noninvasive from invasive carcinoma,^{11,12} to distinguish benign from malignant diseases with overlapping morphologies,^{13–16} to assess estimates of tumor kinetics,¹⁷ to assess reproducibility of grading schemes,¹⁸ to determine whether some subtypes of a disease classification are more easily identified than others,¹⁷ to determine the benefit of additional immunohistochemical (IHC) or molecular testing on morphologic diagnostic accuracy,¹⁹ and to determine whether diagnostic criteria/classification is equally well applied by general and specialist pathologists.^{9,10,18,20} Imperfect IPDA implies that there is room for improvement in diagnostic criteria, diagnostic classifications, and/or pathologist education. Using non-small cell lung carcinoma (NSCLC) classification as a model, this study was designed to test whether addition of a standard panel of mucin and IHC stains to the hematoxylin-eosin (H&E) stain alone has significantly improved IPDA, whether recent NSCLC reclassifications have significantly improved IPDA, and whether the current NSCLC classification using H&E/mucin/IHC stains can be equally well applied independently of practice type, pulmonary pathology expertise, years of experience, and/or lung carcinoma service load.

For pathologic diagnosis of lung carcinomas, prior publications have estimated kappas for the series of evolving lung carcinoma classifications (mainly the World Health Organization [WHO] classifications of 1967,²¹ 1981,²² 1999,²³ 2004,²⁴ and 2015²⁵). Classifications have evolved in parallel with the adoption of novel chemistries, first for detection of intracytoplasmic mucin, then for antibody detection of proteins in paraffin sections that correlated with NSCLC subtype. No kappa data are identified in the literature for the original 1967 WHO classification of lung carcinomas.²¹ The periodic acid-Schiff–diastase (PAS-D) mucin stain was integrated into case diagnostic workup between the 1967 and 1981 WHO classifications. Paech et al²⁶ reviewed 7 studies using the 1967 and 1981 WHO classifications that calculated kappas from 0.48 to 0.88 for the H&E-only dichotomous distinction of squamous from nonsquamous NSCLC. Three groups published kappa data on the 1981 WHO lung carcinoma classification. Campobasso et al²⁷ estimated $k = 0.56–0.87$ for the H&E + mucin stain diagnoses of 639 NSCLC cases by 2 pathologists. Burnett et al²⁸ estimated $k = 0.25$ for the H&E-only NSCLC diagnoses of a set of 100 small cell lung carcinoma (SCLC) and NSCLC biopsy cases by 11 pathologists (1 pulmonary pathology expert), using the 1981 WHO classification.²² Their follow-up study²⁹ estimated $k = 0.39$ for the H&E+mucin stain diagnoses of 100 NSCLC bronchial biopsy cases by 12 pathologists (1

pulmonary pathology expert). Stang et al³⁰ estimated $k = 0.54$ for H&E diagnosis of SCLC and NSCLC, with $k = 0.58$ for diagnosis of squamous carcinoma and $k = 0.55$ for diagnosis of adenocarcinoma. Using the 1999 WHO classification, Colby et al¹² calculated $k = 0.65–0.74$ for the NSCLC diagnoses of a set of 105 SCLC and NSCLC resection cases diagnosed by 4 pulmonary pathology experts. Using the 44-diagnosis 2004 WHO classification, Grilley-Olson et al³¹ calculated $k = 0.25$ for the H&E-only diagnoses of a set of 96 NSCLCs by 24 pathologists, $k = 0.48$ for its 10 major subhead diagnoses, and $k = 0.55$ for the dichotomous squamous/not-squamous classes.

The utility of antibodies such as TTF-1 and cytokeratin (CK) 5/6 as ancillary stains in distinguishing primary lung adenocarcinoma from squamous carcinoma was recognized in the 2004 WHO classification,²⁴ and panels of these IHC stains for this distinction have been developed.^{32–37} Test performance using combinations of a set of mucin and IHC (CK5/6, P63, TTF-1, Napsin A) stains were good for confirmation of morphologically unequivocal cases,³⁷ and for distinction of squamous and adenocarcinoma in morphologically equivocal cases.^{32–36} Using the 2004 WHO classification, Thunnissen et al³⁸ calculated $k = 0.31$ for the H&E-only diagnosis of a set of 37 poorly differentiated NSCLCs by 16 International Association for the Study of Lung Cancer (IASLC) pathologists, with increase to $k = 0.45$ following provision of mucin stains and TTF-1/P63/P40 immunostains. Classification of lung adenocarcinomas was modified by the IASLC/American Thoracic Society/European Respiratory Society (IASLC/ATS/ERS)³⁹ to include micropapillary pattern, to change from diagnosis of “mixed” to “predominant” patterns, to integrate use of IHC stains, and to devise a triage classification for biopsies. These changes were integrated into the 2015 WHO classification of lung carcinomas.²⁵

In the current study, we calculated kappas as a measure of IPDA for H&E-only and H&E/mucin/IHC stain diagnoses of 54 NSCLC cases by 22 pathologists (231 possible pathologist pairs), using 4 current and recent NSCLC diagnostic classifications. These data allowed us to determine how IPDA for H&E-only diagnosis is affected by the addition of a current standard panel of mucin and IHC stains, how IPDA is affected by NSCLC reclassification and simplification, and how IPDA for diagnoses made with a standard panel of H&E/mucin/IHC stains and the current 2015 WHO NSCLC classification is affected by pathologist practice type, practice duration, pulmonary pathology expertise, and routine NSCLC case volume.

MATERIALS AND METHODS

Tested Classifications

Lung carcinoma diagnoses outlined and defined within 3 resection classifications (2004 WHO,²⁴ 2011 IASLC/ATS/ERS,³⁹ and 2015 WHO²⁵) were incorporated into the survey tool. The 2011/2015 triage diagnoses for nonresection (transbronchial biopsy, core biopsy, surgical lung biopsy) specimens^{25,39} were also incorporated into the survey tool. Participants were instructed to review these classifications and their diagnostic criteria before taking the survey.

Institutional Review Board Approval

University of North Carolina (UNC, Chapel Hill) Institutional Review Board approval was obtained before and throughout the performance of these studies (07-1273).

Table 1. Characteristics of the Non–Small Cell Lung Carcinoma Study Pathologists

Participants/No. approached	22/28 (78%)
Prior surgical pathology fellowship	18/22 (82%)
Prior pulmonary pathology fellowship	5/22 (23%)
Current academic practice	15/22 (68%)
Current community practice	7/22 (32%)
No pulmonary pathology expertise	13/22 (59%)
Pulmonary pathology expertise	9/22 (41%)
Years of experience, range	0–39
Lung carcinoma cases diagnosed per year, range	20–800
Lung carcinoma cases as % of total cases, range	1–90

Participants

Anatomic pathologists involved in private or academic practices that include NSCLC cases were approached to participate in the study. The authors sought good representation from pathologists with different practice durations, different lung carcinoma case volumes, different lung carcinoma case percentages, different expertise in pulmonary pathology, and different practice environments (Table 1).

Tissue Microarray, Tissue Sections, and Stains Summary

Paraffin block tissue microarrays (TMAs) had been previously constructed from triplicate 1-mm-diameter core samples of the original UNC “VOILA” (validation of interobserver agreement in lung cancer assessment) set of 96 NSCLC cases.³¹ Five-millimeter-thick TMA sections were stained with H&E, PAS–diastase (PAS–D, mucin), and IHC stains for TTF-1, Napsin A, CK5/6, and P63.

Paraffin Immunohistochemistry Antibodies and Protocols

Napsin A: Tissue microarray sections underwent heat-induced epitope retrieval (HIER)-2 antigen retrieval (cat No. AR9640, pH 9.0; Leica Biosystems, Buffalo Grove, Illinois) × 20 minutes at 100°C. Primary antibody was a mouse monoclonal immunoglobulin (Ig) G2b, clone IP64 (cat. No. NAPSINA-L-CE-S; Novocastra, Leica Biosystems), stock 16 mg/L, working 1:500, binding 15 minutes at room temperature (RT). Bound antibody was detected with Bond Polymer Refine Detection kit (cat. No. DS9800; Leica Biosystems), using 3,3'-diaminobenzidine (DAB) as the chromogen.

TTF-1: Tissue microarray sections underwent HIER-2 antigen retrieval (cat No. AR9640, pH 9.0; Leica Biosystems) × 20 minutes at 100°C. Primary antibody was a mouse monoclonal IgG1 κ, clone SPT24 (cat. No. NCL-TTF-1; Leica Biosystems), stock 75 mg/L, working 1:100, binding 30 minutes at RT. Bound antibody was

detected with Bond Polymer Refine Detection kit, using DAB as the chromogen.

CK5/6: Tissue microarray sections underwent HIER-1 antigen retrieval (cat. no. AR 9961, pH 6.0; Leica Biosystems) × 20 minutes at 100°C. Primary antibodies were mouse monoclonal IgG1, clones D5 and 16B4 (cat. No. 760-4253; Cell Marque/Ventana, Tucson, Arizona), used neat, binding 30 minutes at RT. Bound antibody was detected with Bond Polymer Refine Detection kit, using DAB as the chromogen.

P63: Tissue microarray sections underwent HIER-1 antigen retrieval (cat No. AR 9961, pH 6.0; Leica Biosystems) × 20 minutes at 100°C. Primary antibody was a mouse monoclonal IgG2a/κ, clone 4A4 (cat. No. M7247, DAKO; Agilent, Santa Clara, California), stock 590 mg/L, working 1:150, binding 15 minutes at RT. Bound antibody was detected with Bond Polymer Refine Detection kit, using DAB as the chromogen.

Mucin Staining Protocol

PAS–D (periodic acid–Schiff reaction after diastase digestion) histochemical staining was performed in the UNC histology laboratory. Tissue sections were treated with 1% diastase (α-amylase) in distilled water for 15 minutes at 40°C, rinsed in water for 2 minutes, then placed into 0.5% periodic acid for 12 minutes at 37°C, then rinsed in water, then placed into Schiff solution for 12 minutes at 45°C, rinsed in water for 5 minutes, then counterstained with modified Mayer hematoxylin and bluing reagent, rinsed in tap water, finally dehydrated through alcohols and xylene.

Case Image Selection

All stained sections were scanned at ×20 magnification by using Aperio digital scanning and imaging technology (Aperio, Leica Biosystems). All images were scored for quality. Fifty-four of the original 96 VOILA study cases were selected for use in the current survey, as based on the availability of high-quality matched H&E, mucin, and IHC stain images for each of the cases. Image adequacy for diagnosis was assessed (Table 2). No attempt was made to select for particular quotas of NSCLC subtypes or degrees of differentiation.

Case Image Assembly Into Web Pages

Tagged Image Format (TIF) images were converted to the Deep Zoom Image (DZI) format by using OpenZoom (<https://github.com/openzoom/>; accessed April 1, 2015). These images were uploaded to an Amazon S3 (Simple Storage Service) bucket for remote access. Each of the 54 cases was represented in the survey tool by 2 Web pages. The first of these included a single H&E-stained cross-section of the 1-mm-diameter (0.78 mm²) core biopsy (Figure 1). The second Web page consisted of a 3×2 panel that included the original H&E stain from the first page, but also included the PAS–D (mucin) stain and the 4 immunostains (TTF-1, Napsin A, CK5/6, and P63) (Figure 2). OpenSeadragon ([**Table 2. Non–Small Cell Lung Carcinoma Case Image Quality, Diagnostic Confidence, and Case Grade Distribution**](https://</p>
</div>
<div data-bbox=)

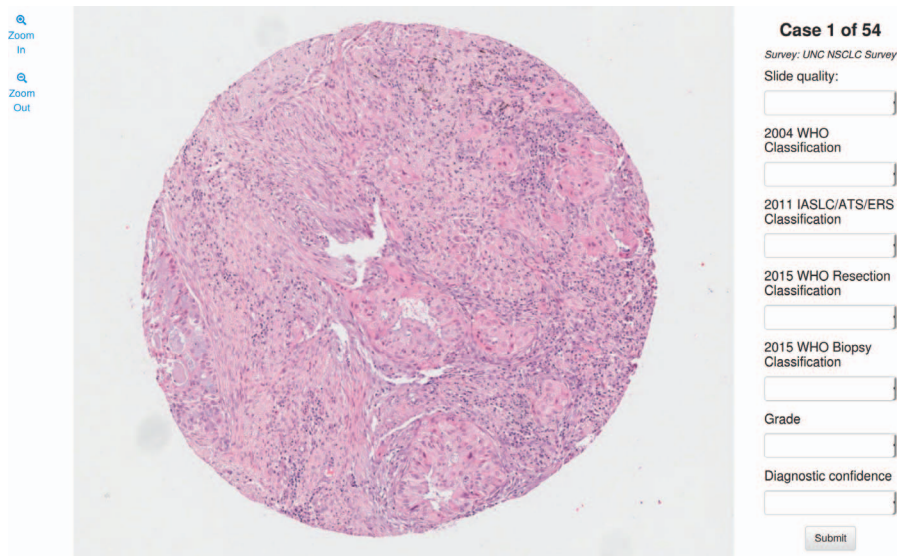
	Interpretation	H&E-only, Frequency (%)	H&E/Mucin/IHC, Frequency (%)	Probability
Image quality	Adequate for diagnosis	1178/1188 (99.2)	1185/1188 (99.7)	<i>P</i> = .03 ^a
Diagnostic confidence	High	326/1188 (27)	811/1188 (68)	<i>P</i> < .001 ^b
	Moderate	592/1188 (50)	305/1188 (26)	
	Low	270/1188 (23)	72/1188 (6)	
Assigned grade	Well differentiated	102/1188 (8)	101/1188 (8)	<i>P</i> = .11 ^b
	Moderately differentiated	469/1188 (39)	456/1188 (38)	
	Poorly differentiated	617/1188 (52)	631/1188 (53)	

Abbreviations: H&E, hematoxylin-eosin stain; H&E/Mucin/IHC, hematoxylin-eosin/periodic acid-Schiff–diastase (PAS–D) mucin/4 immunohistochemical stains (TTF-1, Napsin A, cytokeratin [CK] 5/6, P63).

^a The *P* value shown is for the McNemar test.

^b The *P* value shown is for the test of symmetry (extension of the McNemar test for more than 2 categories) and applies to the entire sets of observations for diagnostic confidence and assigned grade.

Figure 1. Example of first Web page of non-small cell lung carcinoma, case 1: hematoxylin-eosin stain image with survey questions. Abbreviations: IASLC/ATS/ERS, International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society; WHO, World Health Organization.



openseadragon.github.io/; accessed April 1, 2015) was used for display and synchronization of the positioning and zoom level of these images.

Web-Based Survey Tool and Data Collection

A PHP (PHP Hypertext Preprocessor) application was developed and deployed via Heroku (Heroku Co, San Francisco, California). The survey tool contained an initial participant

registration page that assessed demographic information, including surgical pathology fellowship training (Y/N), practice environment (academic/community), pulmonary pathology expertise (Y/N), lung carcinoma service volumes (lung carcinoma cases diagnosed per year), and lung carcinoma percentage (lung carcinoma cases as a percentage of total caseload) (Table 1). Once these demographic questions were answered, the 54 survey cases were opened in order, each case represented by the pairs of Web pages described above, thus 109 Web pages total. Each

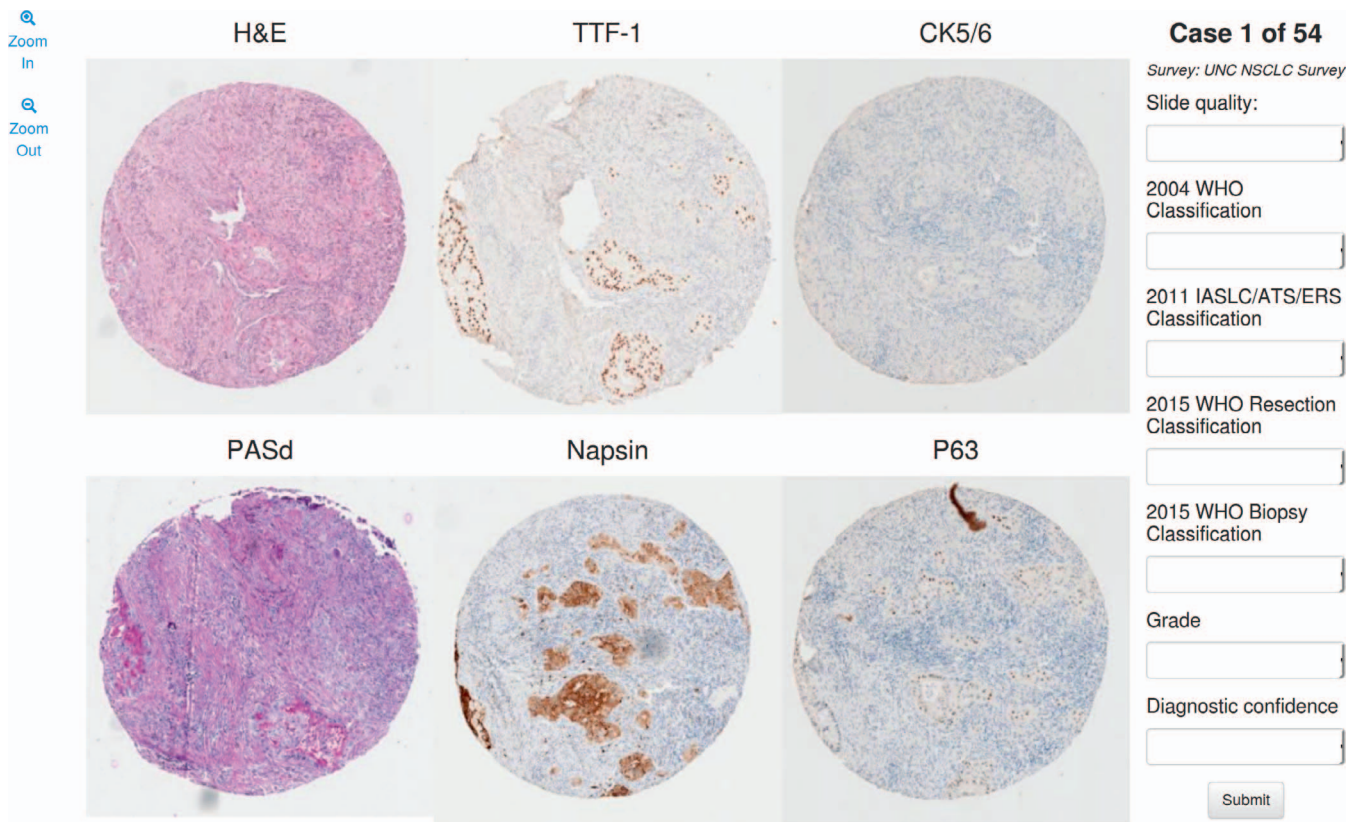


Figure 2. Example of second Web page of non-small cell lung carcinoma, case 1: image panel of hematoxylin-eosin (H&E), periodic acid-Schiff-diastase (PAS-D) mucin, and 4 immunohistochemistry stains (TTF-1, Napsin A, CK5/6, P63), with survey questions. Abbreviations: CK, cytokeratin; IASLC/ATS/ERS, International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society; WHO, World Health Organization.

Table 3. 2004 World Health Organization Lung Tumor Classification^a

Malignant epithelial tumors
Squamous cell carcinoma
Papillary variant
Clear cell variant
Small cell variant
Basaloid variant
Small cell carcinoma
Combined small cell carcinoma
<i>Adenocarcinoma</i>
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 neuroendocrine carcinoma
Combined large cell neuroendocrine carcinoma
Basaloid carcinoma
Lymphoepithelioma-like carcinoma
Clear cell carcinoma
Large cell carcinoma with rhabdoid phenotype
<i>Adenosquamous carcinoma</i>
<i>Sarcomatoid carcinoma</i>
Pleomorphic carcinoma
Spindle cell carcinoma
Giant cell carcinoma
Carcinosarcoma
Pulmonary blastoma
Carcinoid tumor
Typical carcinoid
Atypical carcinoid
<i>Salivary gland tumors</i>
Mucoepidermoid carcinoma
Adenoid cystic carcinoma
Epithelial-myoepithelial carcinoma
Miscellaneous tumors (including mesenchymal tumors and lymphoproliferative tumors)
Metastatic tumors

^a Diagnoses were selectable unless italicized.

pathologist saw the same cases, in the same order. Each Web page for each of the 54 unknown cases contained the same questions (Figures 1 and 2), that is, whether the image was satisfactory for diagnosis (Table 2), the diagnosis of the neoplasm using each of the 4 classifications (Tables 3 through 6), the assigned grade of differentiation (Table 2), and the participant's diagnostic confidence (Table 2). To prevent mucin/IHC stain data from biasing their H&E-only diagnoses, participants were unable to go back and edit their H&E-only diagnoses after seeing the mucin/IHC image grid. Pathologists could exit the survey at any time and

Table 4. 2011 International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society (IASLC/ATS/ERS) Lung Carcinoma Classification^a

<i>Adenocarcinoma</i>
Adenocarcinoma, acinar pattern
Adenocarcinoma, papillary pattern
Adenocarcinoma, micropapillary pattern
Adenocarcinoma, solid with mucin pattern
Adenocarcinoma, lepidic pattern
Mucinous adenocarcinoma with (pattern)
Adenocarcinoma with fetal pattern
Adenocarcinoma with colloid pattern
Adenocarcinoma with (pattern) and signet ring features
Adenocarcinoma with (pattern) and clear cell features
NSCLC, favor adenocarcinoma (morphologic ACa patterns not present, but supported by special stains)
<i>Squamous cell carcinoma</i>
Squamous cell carcinoma
NSCLC, favor squamous carcinoma (morphologic SqCa patterns not present, but supported by special stains)
Small cell carcinoma
Non–small cell carcinoma, NOS
NSCLC with NE morphology (positive NE markers), possible LCNEC
NSCLC with NE morphology (negative NE markers)
<i>Adenosquamous carcinoma</i>
NSCLC with squamous cell and adenocarcinoma patterns, possibly adenosquamous carcinoma
NSCLC, NOS (but immunostains support both squamous and adenocarcinoma components)
<i>Sarcomatoid carcinoma</i>
Poorly differentiated NSCLC with spindle and/or giant cell carcinoma

Abbreviations: ACa, adenocarcinoma; LCNEC, large cell neuroendocrine carcinoma; NE, neuroendocrine; NOS, not otherwise specified; NSCLC, non–small cell lung carcinoma; SqCa, squamous cell carcinoma.

^a Diagnoses were selectable unless italicized.

could restart where they had left off. Demographic and diagnostic results were stored in a PostgreSQL database.

Collapse of Tested Classifications Into Subhead Classes or Dichotomized Classes

Each of the lung tumor classification schemes uses subhead diagnoses, (eg, squamous cell carcinoma, adenocarcinoma) to subgroup the individual diagnoses. The 2004 WHO lung tumor classification²⁴ can be collapsed from 38 individual NSCLC diagnoses into 8 subhead NSCLC diagnoses (see Supplemental Table 1, of 10 supplemental tables, in the Supplemental Digital Content at www.archivesofpathology.org in the December 2018 table of contents). The 2011 IASLC/ATS/ERS NSCLC classification³⁹ can be collapsed from 19 individual NSCLC diagnoses to 7 subhead NSCLC diagnoses (see Supplemental Table 2). The 2015 WHO lung tumor resection classification²⁵ can be collapsed from 32 individual NSCLC diagnoses to 12 subhead NSCLC diagnoses (see Supplemental Table 3). The 2015 WHO lung carcinoma biopsy triage diagnostic algorithm^{25,39} can be collapsed further from 7 individual NSCLC diagnoses to 5 subhead NSCLC diagnoses (see Supplemental Table 4) but is generally already at a subhead diagnosis level of classification. Therapeutically relevant dichotomized classes (squamous/not-squamous [Sq/Not-Sq], adenocarcinoma/not-adenocarcinoma [Ad/Not-Ad]) were derived from these

Table 5. 2015 World Health Organization Lung Tumor Resection Classification^a

<i>Adenocarcinoma</i>
Lepidic adenocarcinoma
Acinar adenocarcinoma
Papillary adenocarcinoma
Micropapillary adenocarcinoma
Solid adenocarcinoma
Invasive mucinous adenocarcinoma
Mixed invasive mucinous and nonmucinous
Colloid adenocarcinoma
Fetal adenocarcinoma
Enteric adenocarcinoma
Minimally invasive adenocarcinoma
Nonmucinous
Mucinous
Preinvasive lesions
Atypical adenomatous hyperplasia
Adenocarcinoma in situ
Nonmucinous
Mucinous
<i>Squamous cell carcinoma</i>
Keratinizing squamous cell carcinoma
Nonkeratinizing squamous cell carcinoma
Basaloid squamous cell carcinoma
Preinvasive lesion
Squamous cell carcinoma in situ
<i>Neuroendocrine tumors</i>
Small cell carcinoma
Combined small cell carcinoma
Large cell neuroendocrine carcinoma
Combined large cell neuroendocrine carcinoma
Carcinoid tumors
Typical carcinoid
Atypical carcinoid
Preinvasive lesion
Diffuse idiopathic neuroendocrine cell hyperplasia
Large cell carcinoma
Adenosquamous carcinoma
Pleomorphic carcinoma
Spindle cell carcinoma
Giant cell carcinoma
Carcinosarcoma
Pulmonary blastoma
Other and unclassified carcinomas
Lymphoepithelioma-like carcinoma
NUT carcinoma
<i>Salivary gland-type tumors</i>
Mucoepidermoid carcinoma
Adenoid cystic carcinoma
Epithelial-myoepithelial carcinoma
Pleomorphic adenoma
<i>Papillomas</i>
Squamous cell papilloma
Exophytic
Inverted
Glandular papilloma
Mixed squamous cell and glandular papilloma

Table 5. Continued

Adenomas
Sclerosing pneumocytoma
Alveolar adenoma
Papillary adenoma
Mucinous cystadenoma
Mucous gland adenoma

Abbreviation: NUT carcinoma, carcinoma associated with “nuclear protein in testis” (*NUTM1*) gene rearrangement.

^a Diagnoses were selectable unless italicized.

subhead categories. The strategy for collapsing the full classifications into subhead classes and dichotomized classes is shown in Supplemental Tables 1 through 4 (of 10 supplemental tables).

Statistical Methods

The Cohen³ kappa was used to measure interpathologist diagnostic agreement among the 231 pathologist pairs from the possible combinations of the 22 study pathologists. Pathologists’ chosen diagnoses for each of the 4 full classification schemes (first with H&E-only, then with H&E plus mucin plus IHC) were also collapsed into subhead classes, and also into dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes (see strategy in Supplemental Tables 1 through 4).⁵ Because these are correlated data, a method known as block bootstrapping^{40,41} was used to calculate group kappa standard errors, through which appropriate 95% CIs were then calculated. Pathologist characteristic subgroups were created by dichotomizing the 22 pathologists into “high” versus “low” groupings. The break point for each of the dichotomizations was chosen at a scientifically relevant point close to or at the median. The characteristics of interest were years of experience, number of lung cancer cases seen in a year, working in an academic or community setting, and whether the pathologist identified as a pulmonary pathologist expert or not. Case image quality, case grade distribution, and diagnostic confidence were compared between the paired diagnoses from H&E-only and then H&E/mucin/IHC by using the McNemar test for 2 categories, or the extension of the McNemar test for 3 categories, the test of symmetry. Analyses were performed with both SAS (version 9.4; SAS Institute, Cary, North Carolina) and R statistical software.⁴²

RESULTS

Participants

The training and practice experiences of the participant pathologists are described in Table 1. Twenty-eight anatomic pathology board-certified pathologists were approached to participate in the study, 26 of these agreed to participate, and 22 of these 26 completed the survey. Of the 22 pathologists, 18 (82%) had completed a surgical pathology fellowship, and 5 (23%) had completed a

Table 6. 2015 World Health Organization Lung Carcinoma Biopsy Triage Classification

Adenocarcinoma
Squamous cell carcinoma
NSCLC, favor adenocarcinoma
NSCLC, favor squamous cell carcinoma
NSCLC, possible adenosquamous carcinoma
NSCLC, ? large cell neuroendocrine carcinoma
NSCLC, NOS
Small cell carcinoma

Abbreviations: NOS, not otherwise specified; NSCLC, non-small cell lung carcinoma.

Table 7. Kappa Variation by Classification, With Estimates for Subhead and Dichotomized Diagnostic Classes^a

Classification	Stated Diagnoses		Collapsed to Subhead Diagnoses		Collapsed to Sq Versus Not-Sq		Collapsed to Ad Versus Not-Ad	
	H&E only	H&E + Muc/IHC	H&E only	H&E + Muc/IHC	H&E only	H&E + Muc/IHC	H&E only	H&E + Muc/IHC
WHO 2004	0.27	0.38 ^b	0.43	0.63 ^b	0.49	0.71 ^b	0.57	0.77 ^b
IASLC/ATS/ERS 2011	0.29	0.42 ^b	0.42	0.74 ^b	0.50	0.83 ^b	0.57	0.84 ^b
WHO 2015 R	0.30	0.49 ^b	0.43	0.74 ^b	0.50	0.84 ^b	0.57	0.83 ^b
WHO 2015 B	0.32	0.45 ^b	0.44	0.75 ^b	0.51	0.85 ^b	0.57	0.83 ^b

Abbreviations: Ad Versus Not-Ad, adenocarcinoma versus not adenocarcinoma; B, biopsy triage; H&E, hematoxylin-eosin stain; H&E + Muc/IHC, hematoxylin-eosin + periodic acid-Schiff–diastase (PAS-D) mucin/4 immunohistochemical stains (TTF-1, Napsin A, cytokeratin [CK] 5/6, P63); IASLC/ATS/ERS, International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society; R, resections; Sq Versus Not-Sq, squamous carcinoma versus not squamous carcinoma; WHO, World Health Organization.

^a Comparison of mean kappas for stated (pathologists') diagnoses, with estimates for mean kappas recalculated after collapse of stated diagnoses into the classifications' major subhead classes or into dichotomized diagnostic classes (see collapse strategies in Supplemental Tables 5 through 8).

^b $P < .05$ (based on nonoverlapping 95% CIs, actual P value not calculated) for the comparison of H&E-only versus H&E/mucin/IHC diagnoses, all pathologists.

pulmonary pathology fellowship. Of the 22 pathologists, 15 (68%) are in academic practice, and 7 (32%) are in community practice. Of the 22 pathologists, 10 (45%) are Pulmonary Pathology Society members, 5 (23%) of whom had participated in the original VOILA study. Of the 22 pathologists, 12 (55%) are not Pulmonary Pathology Society members, 2 (9%) of whom had participated in the original VOILA study. Of the 22 pathologists, 9 (41%) identify as pulmonary pathology experts. Participants had a broad range of practice duration (0–39 years), recurring exposure to lung carcinoma (20–800 new lung carcinoma diagnoses per year), and lung carcinoma–focused practice (1%–90% of cases diagnosed per service week are new lung carcinomas).

Case Image Quality

Participants were asked whether slide images were suitable for diagnosis. Slide images were considered suitable for diagnosis in 2363 of 2376 responses (99.4%) (Table 2).

Diagnostic Confidence

Participants were asked about diagnostic confidence for their H&E-only diagnoses, and then again for their H&E/mucin/IHC diagnoses. Diagnostic confidence improved with the provision of mucin and IHC stains ($P < .001$, McNemar test; Table 2). Diagnostic confidence for H&E/mucin/IHC diagnoses for each of the classifications is shown in Supplemental Tables 5 through 8.

Assigned Neoplastic Grade

Participants were asked to grade each carcinoma at the time of their H&E-only diagnosis, then again at the time of their H&E/mucin/IHC diagnosis. The grade distribution was unchanged after the provision of mucin and IHC stains ($P = .11$, McNemar test; Table 2).

Effect on Kappa Following Provision of Supplemental Mucin and IHC Stains

IPDA significantly improved when pathologists were provided with mucin and IHC stains in addition to H&E stains, compared to IPDA when using H&E alone (Table 7). This was true across each of the 4 classifications, as well as following progressive simplification into subhead and dichotomized diagnostic classes (Table 7). For the 2004 WHO, 2011 IASLC/ATS/ERS, and 2015 WHO resection classifications, this IPDA improvement was present across

practice environments, pulmonary pathology expertise, practice duration, and lung carcinoma exposure in service caseload (see Supplemental Table 9). For the 2015 WHO biopsy triage classification, IPDA improved after provision of mucin and IHC stains for participants with less than 6 years in practice, more than 100 new lung carcinoma cases per year, less than 4% lung carcinoma cases in their service caseloads, and for those without pulmonary pathology expertise (see Supplemental Table 9).

Variation in Kappa by NSCLC Classification

Kappas of stated H&E-only NSCLC diagnoses were not significantly different between the 4 classifications (Figure 3; Table 7). Kappas of stated H&E/mucin/IHC NSCLC diagnoses by all current study pathologists showed significant improvement in IPDA for the 2015 WHO resection classification (mean $k = 0.49$ [95% CI, 0.46–0.51]) when compared with IPDA for the 2004 WHO classification (mean $k = 0.38$ [95% CI, 0.36–0.40]) or for the 2011 IASLC/ATS/ERS classification (mean $k = 0.42$ [95% CI, 0.39–0.44]) (Figure 3; Table 7). Kappas of stated H&E/mucin/IHC NSCLC diagnoses by all current study pathologists showed no significant difference in IPDAs between the 2015 WHO resection classification (mean $k = 0.49$ [95% CI, 0.46–0.51]) and the 2015 WHO biopsy triage classification (mean $k = 0.45$ [95% CI, 0.40–0.50]) (Figure 3; Table 7). However, pathologist subgroup analysis shows higher kappas for stated H&E/mucin/IHC diagnoses when using the 2015 WHO resection classification than the 2015 WHO biopsy triage classification for pathologists who diagnose more than 100 cases of lung carcinoma per year, and for those with pulmonary pathology expertise (see Supplemental Table 10).

Variation in Kappa Following Simplification of Classifications Into Subhead or Dichotomized Classes

Table 7 shows mean kappas (all pathologists, all 4 classifications) calculated from the stated H&E-only or H&E/mucin/IHC diagnoses, and includes estimated kappas recalculated after collapse of the stated diagnoses into either subhead classes or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes. Kappas increased after collapse of stated H&E-only or H&E/mucin/IHC diagnoses into subhead or dichotomized classes. Dichotomization (both Sq/Not-Sq and Ad/Not-Ad) estimates kappas above 0.8 for the 2015 WHO

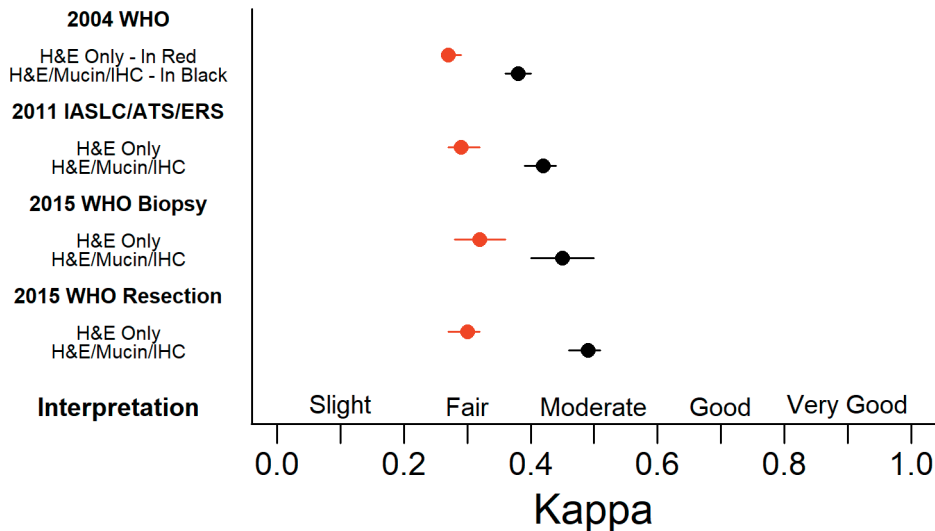


Figure 3. Kappa variation by classification and stains provided. Distributions represent the kappas for diagnoses of all participants. Bars represent the 95% CI around the mean kappa for each comparison. Red represents diagnoses made with hematoxylin-eosin (H&E) stain only. Black represents diagnoses made with H&E, periodic acid-Schiff–diastase (PAS-D) mucin, and 4 immunohistochemical (IHC) stains (TTF-1, Napsin A, CK5/6, P63). Abbreviations: CK, cytokeratin; IASLC/ATS/ERS, International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society; WHO, World Health Organization.

resection and biopsy triage classifications (all pathologists, H&E/mucin/IHC stains). Table 8 shows kappa variation by pathologist subgroup for H&E/mucin/IHC diagnoses, using the WHO 2015 resection classification, and includes estimated kappas recalculated after collapse of stated diagnoses into subhead or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes. Dichotomization (both Sq/Not-Sq and Ad/Not-Ad) estimates $k \geq 0.8$ for each pathologist subgroup when using H&E/mucin/IHC stains and the 2015 WHO resection classification.

Variation in Kappa by Practice Type

Of the 22 pathologists, 15 (68%) are in academic practices, and 7 (32%) are in community practices. Using the current 2015 WHO resection classification, IPDA for diagnoses using H&E/mucin/IHC stains was not significantly different between academic (mean $k = 0.49$ [95% CI, 0.46–0.52]) and community (mean $k = 0.45$ [95% CI, 0.39–0.52]) pathologists (Figure 4; Table 8). Significant improvement is estimated for kappas recalculated after collapse into subhead or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes for each of these 2

Table 8. Kappa Variation by Pathologist Subgroup, 2015 World Health Organization Lung Tumor Resection Classification^{a-c}

	Kappa (95% CI) for Diagnoses Based on H&E/Mucin/IHC Stains			
	Stated Diagnoses	Collapsed to Subhead Diagnoses	Collapsed to Sq Versus Not-Sq	Collapsed to Ad Versus Not-Ad
Academic	0.49 (0.46–0.52)	0.74 (0.71–0.77)	0.83 (0.81–0.85)	0.84 (0.81–0.87)
Community	0.45 (0.39–0.52)	0.74 (0.70–0.78)	0.86 (0.84–0.87)	0.81 (0.76–0.86)
Pulm path exp	0.62 (0.59–0.65) ^d	0.80 (0.78–0.82) ^d	0.87 (0.85–0.90) ^d	0.89 (0.88–0.91) ^d
Not pulm path exp	0.42 (0.40–0.44) ^d	0.70 (0.68–0.73) ^d	0.82 (0.80–0.84) ^d	0.79 (0.76–0.82) ^d
<6 y in practice	0.45 (0.42–0.48) ^d	0.71 (0.67–0.74) ^d	0.80 (0.78–0.82) ^d	0.80 (0.76–0.84)
>6 y in practice	0.54 (0.48–0.60) ^d	0.78 (0.74–0.82) ^d	0.88 (0.86–0.90) ^d	0.86 (0.81–0.90)
<100 Lung Ca/y	0.43 (0.40–0.46) ^d	0.72 (0.69–0.75) ^d	0.84 (0.83–0.86)	0.78 (0.75–0.82) ^d
>100 Lung Ca/y	0.58 (0.55–0.60) ^d	0.77 (0.76–0.79) ^d	0.84 (0.81–0.87)	0.89 (0.87–0.90) ^d
<4% Lung Ca	0.46 (0.41–0.50)	0.73 (0.70–0.77)	0.86 (0.84–0.88)	0.80 (0.75–0.84)
>4% Lung Ca	0.52 (0.49–0.56)	0.74 (0.72–0.77)	0.82 (0.79–0.84)	0.86 (0.83–0.88)

Abbreviations: Academic, academic practice; Ad Versus Not-Ad, adenocarcinoma versus not adenocarcinoma; Community, community practice; H&E/Mucin/IHC, hematoxylin-eosin/periodic acid-Schiff–diastase (PAS-D) mucin/4 immunohistochemistry stains (TTF-1, Napsin A, cytokeratin [CK] 5/6, P63); Not pulm path exp, not a pulmonary pathology expert; Pulm path exp, pulmonary pathology expert; Sq Versus Not-Sq, squamous carcinoma versus not squamous carcinoma; >4% Lung Ca, lung carcinoma diagnoses comprise more than 4% of the participant pathologist's service cases; >100 Lung Ca/y, participant pathologist diagnoses more than 100 new lung carcinomas per year.

^a Comparison of kappa distributions (mean 95% CI) for stated (pathologists') diagnoses, with estimates for kappa distributions recalculated after collapse of stated diagnoses into the classifications' major subhead classes or into dichotomized diagnostic classes (see collapse strategies in Supplemental Tables 5 through 8).

^b Kappas are based on diagnoses that used H&E/Mucin/IHC stains and used the 2015 WHO lung tumor resection classification.

^c Cut points represent the medians across the participants for numbers and percentages of lung carcinoma cases diagnosed when on service.

^d $P < .05$ (based on nonoverlapping 95% CIs, actual P value not calculated) for comparisons of the different subgroups (practice type, pulmonary pathology expertise, years in practice, lung carcinoma cases diagnosed per year, and lung carcinoma cases as a percentage of total caseload).

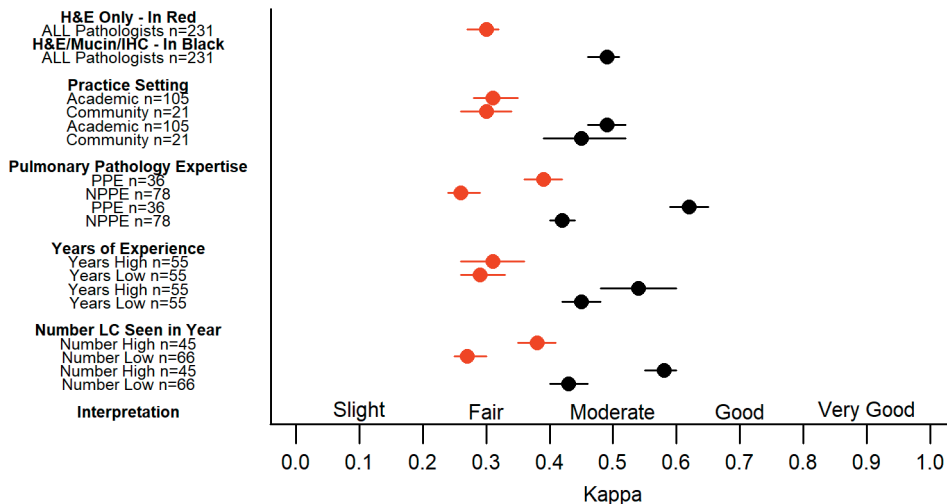


Figure 4. Kappa variation by pathologist characteristics and stains provided. Distributions represent the kappas for diagnoses made when using the 2015 World Health Organization resection classification. Bars represent the 95% CI around the mean kappa for each comparison. Red represents diagnoses made with hematoxylin-eosin (H&E) stain only. Black represents diagnoses made with H&E, periodic acid-Schiff–diastase (PAS-D) mucin, and 4 immunohistochemical (IHC) stains (TTF-1, Napsin A, CK5/6, P63). “n”: number of pathologist pairs compared; Years of Experience, High: more than 6 years in practice; Number LC Seen in Year, High: more than 100 new lung carcinomas diagnosed per year. Abbreviations: CK, cytokeratin; LC, lung carcinomas; NPPE: not a pulmonary pathology expert; PPE, pulmonary pathology expert.

groups, without significant differences between the 2 groups (Table 8).

Variation in Kappa by Pulmonary Pathology Expertise

Of the 22 pathologists, 9 (41%) identified as pulmonary pathology experts, and 13 of 22 participants (59%) did not identify as pulmonary pathology experts. Using the current 2015 WHO resection classification, IPDA for stated diagnoses using H&E/mucin/IHC stains was significantly higher for pathologists identifying as pulmonary pathology experts (mean $k = 0.62$ [95% CI, 0.59–0.65]) than for pathologists not identifying as pulmonary pathology experts (mean $k = 0.42$ [95% CI, 0.40–0.44]) (Figure 4; Table 8). Significant improvement is estimated for kappas recalculated after collapse into subhead or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes for each of these 2 groups, with significant differences between the 2 groups (Table 8).

Variation in Kappa by Years in Practice

Participant practice duration ranged from 0 to 39 years, with a median of 6 years. Using the current 2015 WHO resection classification, IPDA for diagnoses using H&E/mucin/IHC stains was significantly higher for pathologists with more than 6 years of practice duration (mean $k = 0.54$ [95% CI, 0.48–0.60]) than for those with 6 or fewer years of practice duration (mean $k = 0.45$ [95% CI, 0.42–0.48]) (Figure 4; Table 8). Significant improvement is estimated for kappas recalculated after collapse into subhead or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes for each of these 2 groups, with significant differences between the 2 groups for subhead and Sq/Not-Sq classes (Table 8).

Variation in Kappa by Number of New Lung Carcinomas Diagnosed per Year

Participant practice exposure to new lung carcinomas each year ranged from 20 to 800 new lung carcinoma cases per year, with a median of 100 cases per year. Using the current 2015 WHO resection classification, IPDA for diagnoses

using H&E/mucin/IHC was significantly higher for pathologists who see more than 100 new lung carcinoma cases per year (mean $k = 0.58$ [95% CI, 0.55–0.60]) than for those who see fewer than 100 new lung carcinoma cases per year (mean $k = 0.43$ [95% CI, 0.40–0.46]) (Figure 4; Table 8). Significant improvement is estimated for kappas recalculated after collapse into subhead or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes for each of these 2 groups, with significant differences between the 2 groups for subhead and Ad/Not-Ad classes (Table 8).

Variation in Kappa by Percentage of Cases per Service Week That Are Lung Carcinomas

Participant practice exposure to lung carcinoma as a percentage of total new cases ranged from 1% to 90%, with a median of 4%. Using the current 2015 WHO resection classification, IPDA for diagnoses using H&E/mucin/IHC is not significantly different for pathologists who see greater than 4% new lung carcinoma cases per total caseload (mean $k = 0.52$ [95% CI, 0.49–0.56]) than for those who see 4% or fewer (mean $k = 0.46$ [95% CI, 0.41–0.50]) (Table 8). Significant improvement is estimated for kappas recalculated after collapse into subhead or dichotomized (Sq/Not-Sq, Ad/Not-Ad) classes for each of these 2 groups, without significant differences between the 2 groups (Table 8).

DISCUSSION

The current study’s Web-based survey tool allowed unbiased H&E-only diagnosis, followed immediately by H&E/mucin/IHC diagnosis, of a series of 54 NSCLC cases by 22 pathologists. Cohen’s kappa was calculated (231 pathologist-pairs) as the measure of IPDA. This study allowed comparison of IPDA for H&E-only and H&E/mucin/IHC diagnoses across 4 NSCLC classifications, allowed estimation of IPDA following collapse of full classifications into subhead and therapeutically relevant dichotomized classes, and allowed comparison of IPDA for pathologist subgroups that differ by practice type, pulmo-

nary pathology expertise, years in practice, and lung carcinoma service load.

IPDA for H&E-Only Diagnoses of NSCLC

The current study allowed comparison of H&E-only diagnoses of NSCLC using the WHO 2004 classification with our previous VOILA study data.³¹ The current 54 cases are a subset of the 96-case VOILA set. Only 7 of the current 22 pathologists participated in both studies. The VOILA study presented whole sections, whereas the current study presented 1-mm-diameter cores. We found that kappas for H&E-only diagnoses of NSCLC are similar between our previous study (mean $k = 0.25$ [95% CI, 0.23, 0.26]) and our current study (mean $k = 0.27$ [95% CI, 0.26, 0.29]). These H&E-only data suggest that the current 54-case subset is representative of the original 96-case VOILA set, that kappa is not affected by changes in the set of diagnostic pathologists, that kappa is not affected by the reduced square area of a 1-mm core cross-section, and that the same diagnostic criteria can be applied to both whole-section and 1-mm core biopsies.

Burnett et al²⁸ estimated mean $k = 0.25$ for the H&E-only diagnoses of the NSCLCs in a set of 100 SCLC and NSCLC biopsy cases by 11 pathologists (1 pulmonary pathology expert) using the 1981 WHO classification.²² A recent study by Thunnissen et al³⁸ of 37 cases of resected poorly differentiated NSCLC diagnosed by 16 IASLC Pathology Committee pathologists using the 2004 WHO morphologic criteria found mean $k = 0.31$ (95% CI, 0.23, 0.40) for H&E-only diagnosis. Overall, these H&E-only data estimate IPDA for H&E-only diagnosis of NSCLC at mean $k = 0.25$ – 0.35 . Our current H&E-only kappas for the subsequent 2011 and 2015 classifications are similar. Although not the current standard of practice (which uses ancillary mucin and IHC stains), these H&E-only data provide a baseline for testing whether the addition of a standard panel of mucin and IHC stains can improve IPDA for NSCLC.

IPDA for H&E/Mucin/IHC Diagnoses of NSCLC

We found that provision of mucin and IHC stains to participant pathologists led to significant improvement in IPDA for all 4 NSCLC classifications studied. For the current 2015 WHO resection classification, overall kappas (all 22 current-study pathologists, all cases) improved from mean $k = 0.32$ (H&E-only) to mean $k = 0.49$ (H&E/mucin/IHC). For the current 2015 WHO resection classification, this improvement was found across practice environments, across pulmonary pathology expertise, across practice duration, and across lung carcinoma exposure in service caseload.

Burnett et al²⁹ found that addition of mucin stains to H&E-alone diagnoses achieved an estimated mean $k = 0.39$ for the diagnosis of 100 NSCLC bronchial biopsy cases by 12 pathologists (1 pulmonary pathology expert) using the 1981 WHO classification.²² It is not clear whether this set of NSCLC cases is the same or different from that of their 1994 study,²⁸ so it is unclear whether the change in kappa from mean $k = 0.25$ to mean $k = 0.39$ following addition of mucin stains is comparable. The more recent study by Thunnissen et al³⁸ of 37 cases of resected poorly differentiated NSCLC diagnosed by 16 IASLC Pathology Committee pathologists using the 2004 WHO morphologic criteria found that overall kappa increased from mean $k = 0.31$ (95% CI, 0.23, 0.40) with H&E alone to mean $k = 0.45$ (95% CI, 0.37, 0.53) after provision of mucin, TTF-1, and P63/P40 stains³⁸; solid pattern adenocarcinoma improved from mean $k = 0.21$ to

mean $k = 0.63$ after provision of these stains. These data support the use of supplemental mucin and immunohistochemical stains as current standard of practice for any case that is not confidently diagnosed by H&E alone. Use of a limited IHC panel appears to be beneficial in cytology cell blocks or small biopsy samples, with 85% to 100% prediction of diagnosis in the subsequent resection.^{33,36} Use of a limited IHC panel may also improve diagnostic accuracy, given published estimates of an estimated 6% change in diagnosis of squamous cell carcinomas following IHC testing.⁴³

Differences in IPDA Between Current and Recent NSCLC Classifications

We found that overall kappas (all 22 current-study pathologists, all cases, H&E/mucin/IHC stains) improved between the 2004 WHO classification and the 2015 WHO classification (see Table 7). Current study mean kappa using H&E/mucin/IHC stains has improved significantly from mean $k = 0.38$ (95% CI, 0.36, 0.40) for the 2004 WHO classification to mean $k = 0.42$ (95% CI, 0.39, 0.44) for the 2011 IASLC classification, to mean $k = 0.49$ (95% CI, 0.46, 0.51) for the 2015 WHO resection classification, and to mean $k = 0.45$ (95% CI, 0.40, 0.50) for the 2015 WHO biopsy triage classification. We have not identified other IPDA studies in the literature directly comparing the different recent NSCLC classifications. Our current data suggest that reclassification and refinement of diagnostic criteria, including reference to the value of mucin and a defined panel of immunohistochemical stains to supplement H&E-only diagnosis, have improved IPDA for NSCLC.

Estimated IPDA After Simplification Into Subhead and Dichotomized Classes

Current study data show that kappas for both H&E-only diagnoses and H&E/mucin/IHC diagnoses improve with collapse of stated diagnoses into subhead and dichotomized classes. The caveat here is that only the stated diagnoses based on the full classifications are data entered by the participant pathologists; the estimated kappas for subhead and dichotomized classes are calculated data derived after binning of the stated diagnoses into the collapsed classes (see collapse strategy in Supplemental Tables 1 through 4). For H&E/mucin/IHC diagnoses (all pathologists) using the 2015 WHO resection classification, kappas are estimated to improve from mean $k = 0.49$ (for the stated diagnoses) to mean $k = 0.74$ (for the 10 major subhead diagnose), and to mean $k = 0.84$ (Sq/Not-Sq) or mean $k = 0.83$ (Ad/Not-Ad) for these therapeutically relevant dichotomized classes. We did not identify comparable literature for evaluation of calculated IPDA following collapse of specific diagnoses into major subhead or dichotomized classes.

Differences in IPDA by Practice Environment, Pulmonary Pathology Expertise, Years in Practice, and Practice Exposure to Lung Carcinomas

We found no significant difference in IPDA between participant pathologists from academic and community practice environments (all classifications, H&E-only or H&E/mucin/IHC diagnoses). These data support the argument that standardized NSCLC classification schemes and diagnostic criteria are being defined, published, and distributed effectively across different practice environments. We found higher IPDA for pathologists with

pulmonary pathology expertise, for pathologists with more than 6 years in practice, and for pathologists who diagnose more than 100 new lung carcinomas per year. These data are similar to those seen in our previous H&E-only study, which found multivariate significant difference by pulmonary pathology expertise.³¹ Relative importance of different diagnostic morphologic criteria for poorly differentiated NSCLC can be difficult to define,³⁸ so improved IPDA might be expected for pathologists familiar with complex, occasionally overlapping, diagnostic criteria, and for those who see larger numbers of NSCLC cases when on service.

Potential Limitations of the Study

Potential limitations include the limited set of cases for diagnosis, the limited tissue for evaluation, the lack of usual clinical and radiographic data, the use of digital tools for diagnosis, the use of an imperfect IHC test panel, the use of the median as a test break point for practice variable comparisons, and the request of participant pathologists to apply diagnostic criteria from the 4 different classifications. The set of 54 cases will be limited in its representation of uncommon types of NSCLC, an unavoidable consequence of an unselected case series. One-millimeter-diameter (0.78 mm²) core square area is less than the square area of a typical 1 to 4 cm² (100–400 mm²) tissue section, and does not mimic the fragmented endobronchial and transbronchial biopsies typical in diagnostic practice. Our goal was to test the reproducibility of the classifications, not to propose or test whether 1-mm core sections are equivalent to resection whole sections or biopsy sections. However, the comparable H&E-only/2004 WHO kappa distributions observed with the prior VOILA (whole section) study and the current (1-mm core) study support the argument that representative section square area does not affect the use of the 2004 WHO resection classification criteria for calculation of kappa. All of the current survey cases were invasive lung primaries, but unselected small biopsy sections would limit a pathologist's ability to make diagnoses that are contingent on size (minimally invasive adenocarcinoma) or identifiable subsets (eg, large cell carcinoma, adenosquamous carcinoma). Thus, it is conceivable that reduced square area could affect kappas when using the diagnostic criteria for the other classifications, a possibility that cannot be answered with our current dataset. The absence of clinical and radiographic data allows a cleaner study design, but requires the participant's assumption that all cases are lung primaries. Digital images show comparable IPDA to glass slides⁴⁴ and offer ease-of-use for large numbers of pathologists to diagnose large numbers of cases. The test performance of the panel used for this study has a well-described limitation, that is, the nonspecificity of P63 for squamous carcinoma^{33,36}; fortunately, the complementary stains for adenocarcinoma are more specific, such that coexpression supports adenocarcinoma. Use of the median as a test break point for practice variables was arbitrary, but intentional, so as to lead to similar number of pairs in each subgroup for kappa calculation. Participants were instructed to review the diagnostic criteria for each of the 4 classification systems before starting the survey, but keeping the 4 classifications' differences in mind was a challenge, and may have contributed to some diagnostic variation between participants.

Potential Implications of the Study

To the treating clinician, reproducible pathologic diagnoses and disease classifications are critical to decisions regarding patient management. It is our opinion that routine measurement of IPDA will facilitate empirical improvement of diagnostic criteria and disease classifications, so as to make our pathologic diagnoses more reproducible, more accurate, and thus more trustworthy. More reproducible pathologic diagnoses should lead to more accurate trial arm assignments, more appropriate treatments, and better clinical outcomes. It follows that practicing pathologists will benefit by publication and use of demonstrably reproducible classifications, by adherence to published diagnostic criteria, by participation in continuing medical education courses that teach updated classification systems, and by recognition of diagnoses for which there is poor IPDA. Because most diagnoses are rendered by a single pathologist without further review, it is incumbent upon diagnostic pathologists to be cognizant of those diagnoses and classifications that are inherently nonreproducible, thus meriting internal or external review before case finalization. As a discipline, diagnostic pathologists need to continue to improve IPDA through use of clear diagnostic criteria crafted into nonoverlapping disease classifications, and to educate pathologists at all stages of practice about use of these disease classifications. Diagnoses found to be nonreproducible should be further examined regarding whether the diagnostic criteria are insufficient, poorly defined, nonstandardized, or overlapping, and should be modified until reproducible. It follows that new classification schemes should be validated for maximal IPDA before distribution.

CONCLUSIONS

In this study, IPDA is significantly higher for H&E/mucin/IHC diagnoses than for H&E-only diagnoses. IPDA for H&E/mucin/IHC diagnoses is significantly higher for the 2015 NSCLC classification than for the 2004 or 2011 classifications. IPDA for H&E/mucin/IHC diagnoses using the 2015 NSCLC classification is significantly higher for pathologists who identify as pulmonary pathology experts, have more years of experience, and/or see a high number of lung cancer cases. IPDA for H&E-only or H&E/mucin/IHC diagnoses is similar for community and academic pathologists, supporting the argument that NSCLC diagnostic criteria are being clearly defined, broadly published, read, and applied by practicing pathologists.

Research was partially supported by a grant from the Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill. We thank Michele C. Hayward, RD, for administrative support.

References

1. Kohn LT, Corrigan J, Donaldson MS. *To Err is Human: Building a Safer Health System*. Washington, DC: National Academy Press; 2000.
2. Raab SS, Grzybicki DM, Janosky JE, et al. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. *Cancer*. 2005;104(10):2205–2213.
3. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
4. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213–220.
5. Fleiss JL. Measuring nominal scale agreement among many raters *Psychol Bull*. 1971;76(5):378–381.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
7. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.

8. Uebersax JS. A design-independent method for measuring the reliability of psychiatric diagnosis. *J Psychiatr Res*. 1982;17(4):335–342.
9. Kerkhof M, van Dekken H, Steyerberg EW, et al. Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*. 2007;50(7):920–927.
10. Carlson JW, Jarboe EA, Kindelberger D, Nucci MR, Hirsch MS, Crum CP. Serous tubal intraepithelial carcinoma: diagnostic reproducibility and its implications. *Int J Gynecol Pathol*. 2010;29(4):310–314.
11. Anderson TJ, Sufi F, Ellis IO, Sloane JP, Moss S. Implications of pathologist concordance for breast cancer assessments in mammography screening from age 40 years. *Hum Pathol*. 2002;33(3):365–371.
12. 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.
13. Baumann I, Fuhrer M, Behrendt S, et al. Morphological differentiation of severe aplastic anaemia from hypocellular refractory cytopenia of childhood: reproducibility of histopathological diagnostic criteria. *Histopathology*. 2012;61(1):10–17.
14. Farmer ER, Gonin R, Hanna MP. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. *Hum Pathol*. 1996;27(6):528–531.
15. Fadare O, Parkash V, Dupont WD, et al. The diagnosis of endometrial carcinomas with clear cells by gynecologic pathologists: an assessment of interobserver variability and associated morphologic features. *Am J Surg Pathol*. 2012;36(8):1107–1118.
16. Gerami P, Busam K, Cochran A, et al. Histomorphologic assessment and interobserver diagnostic reproducibility of atypical spitzoid melanocytic neoplasms with long-term follow-up. *Am J Surg Pathol*. 2014;38(7):934–940.
17. Hasegawa T, Yamamoto S, Nojima T, et al. Validity and reproducibility of histologic diagnosis and grading for adult soft-tissue sarcomas. *Hum Pathol*. 2012;43(1):111–115.
18. Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JL. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol*. 2001;32(1):81–88.
19. Vang R, Gupta M, Wu LS, et al. Diagnostic reproducibility of hydatidiform moles: ancillary techniques (p57 immunohistochemistry and molecular genotyping) improve morphologic diagnosis. *Am J Surg Pathol*. 2012;36(3):443–453.
20. Allsbrook WC Jr, Mangold KA, Johnson MH, et al. Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. *Hum Pathol*. 2001;32(1):74–80.
21. Kreyberg L, Liebow AA, Uehlinger EA. *Histologic Typing of Lung Tumors*. 1st ed. Geneva: World Health Organization; 1967.
22. World Health Organization. *Histological Typing of Lung Tumours*. 2nd ed. Geneva: World Health Organization; 1981.
23. Travis WD, Colby TV, Corrin B, Shimosato Y, Brambilla E. *Histological Typing of Lung and Pleural Tumors*. 3rd ed. New York: Springer; 1999.
24. Travis WD, Brambilla E, Muller-Hermelink HK, Harris CC. *Pathology and Genetics of Tumours of the Lung, Pleura, Thymus, and Heart*. 3rd ed. Lyon, France: IARC; 2004. *World Health Organization Classification of Tumours*; vol 10.
25. Travis WD, Brambilla C, Burke AP, Marx A, Nicholson AG. *WHO Classification Tumours of the Lung, Pleura, Thymus, and Heart*. Lyon, France: IARC; 2015. *World Health Organization Classification of Tumours*; vol 7.
26. Paech DC, Weston AR, Pavlakis N, et al. A systematic review of the interobserver variability for histology in the differentiation between squamous and nonsquamous non-small cell lung cancer. *J Thorac Oncol*. 2011;6(1):55–63.
27. Campobasso O, Andron 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.
28. 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.
29. 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.
30. Stang A, Pohlabein 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.
31. Grilley-Olson JE, Hayes DN, Moore DT, et al. Validation of interobserver agreement in lung cancer assessment: hematoxylin-eosin diagnostic reproducibility for non-small cell lung cancer: the 2004 World Health Organization classification and therapeutically relevant subsets. *Arch Pathol Lab Med*. 2013;137(1):32–40.
32. Whithaus K, Fukuoka J, Prihoda TJ, Jagirdar J. Evaluation of napsin A, cytokeratin 5/6, p63, and thyroid transcription factor 1 in adenocarcinoma versus squamous cell carcinoma of the lung. *Arch Pathol Lab Med*. 2012;136(2):155–162.
33. Rekhtman N, Ang DC, Sima CS, Travis WD, Moreira AL. Immunohistochemical algorithm for differentiation of lung adenocarcinoma and squamous cell carcinoma based on large series of whole-tissue sections with validation in small specimens. *Mod Pathol*. 2011;24(10):1348–1359.
34. Nicholson AG, Gonzalez D, Shah P, et al. Refining the diagnosis and EGFR status of non-small cell lung carcinoma in biopsy and cytologic material, using a panel of mucin staining, TTF-1, cytokeratin 5/6, and P63, and EGFR mutation analysis. *J Thorac Oncol*. 2010;5(4):436–441.
35. Loo PS, Thomas SC, Nicolson MC, Fyfe MN, Kerr KM. Subtyping of undifferentiated non-small cell carcinomas in bronchial biopsy specimens. *J Thorac Oncol*. 2010;5(4):442–447.
36. Mukhopadhyay S, Katzenstein AL. Subclassification of non-small cell lung carcinomas lacking morphologic differentiation on biopsy specimens: utility of an immunohistochemical panel containing TTF-1, napsin A, p63, and CK5/6. *Am J Surg Pathol*. 2011;35(1):15–25.
37. Kim MJ, Shin HC, Shin KC, Ro JY. Best immunohistochemical panel in distinguishing adenocarcinoma from squamous cell carcinoma of lung: tissue microarray assay in resected lung cancer specimens. *Ann Diagn Pathol*. 2013;17(1):85–90.
38. Thunnissen E, Noguchi M, Aisner S, et al. Reproducibility of histopathological diagnosis in poorly differentiated NSCLC: an international multiobserver study. *J Thorac Oncol*. 2014;9(9):1354–1362.
39. 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.
40. Liu RJ, Singh K. Using i.i.d. bootstrap inference for general non-i.i.d. models. *J Stat Plan Infer*. 1995;43:67–75.
41. DasGupta A, ed. *SpringerLink: Asymptotic Theory of Statistics and Probability*. New York, NY: Springer; 2008. Springer Texts in Statistics.
42. R: *A Language and Environment for Statistical Computing* [computer program]. Vienna, Austria: R foundation for statistical computing; 2008.
43. Kadota K, Nitadori J, Rekhtman N, Jones DR, Adusumilli PS, Travis WD. Reevaluation and reclassification of resected lung carcinomas originally diagnosed as squamous cell carcinoma using immunohistochemical analysis. *Am J Surg Pathol*. 2015;39(9):1170–1180.
44. Ozluk Y, Blanco PL, Mengel M, Solez K, Halloran PF, Sis B. Superiority of virtual microscopy versus light microscopy in transplantation pathology. *Clin Transplant*. 2012;26(2):336–344.