Annals of Diagnostic Pathology 47 (2020) 151561

Contents lists available at ScienceDirect

Annals of Diagnostic Pathology

journal homepage: www.elsevier.com/locate/anndiagpath

Evidence Based Pathology

Pathologists should probably forget about kappa. Percent agreement, diagnostic specificity and related metrics provide more clinically applicable measures of interobserver variability

Alberto M. Marchevsky[a,*], Ann E. Walts[a], Birgit I. Lissenberg-Witte[b], Erik Thunnissen[c]

[a] Department of Pathology & Laboratory Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States of America
[b] Department of Epidemiology and Data Science, UMC, Vrije Universiteit Amsterdam, the Netherlands
[c] Department of Pathology, UMC, Vrije Universiteit Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Diagnostic accuracy
Kappa statistics
Interobserver variability
Evidence-based pathology

ABSTRACT

Kappa statistics have been widely used in the pathology literature to compare interobserver diagnostic variability (IOV) among different pathologists but there has been limited discussion about the clinical significance of kappa scores. Five representative and recent pathology papers were queried using clinically relevant specific questions to learn how IOV was evaluated and how the clinical applicability of results was interpreted. The papers supported our anecdotal impression that pathologists usually assess IOV using Cohen's or Fleiss' kappa statistics and interpret the results using some variation of the scale proposed by Landis and Koch. The papers did not cite or propose specific guidelines to comment on the clinical applicability of results. The solutions proposed to decrease IOV included the development of better diagnostic criteria and additional educational efforts, but the possibility that the entities themselves represented a continuum of morphologic findings rather than distinct diagnostic categories was not considered in any of the studies.

A dataset from a previous study of IOV reported by Thunnissen et al. was recalculated to estimate percent agreement among 19 international lung pathologists for the diagnosis of 74 challenging lung neuroendocrine neoplasms. Kappa scores and diagnostic sensitivity, specificity, positive and negative predictive values were calculated using the majority consensus diagnosis for each case as the gold reference diagnosis for that case. Diagnostic specificity estimates among multiple pathologists were > 90%, although kappa scores were considerably more variable. We explain why kappa scores are of limited clinical applicability in pathology and propose the use of positive and negative percent agreement and diagnostic specificity against a gold reference diagnosis to evaluate IOV among two and multiple raters, respectively.

1. Introduction

The results and conclusions of many studies in Pathology are considered clinically valid only after significant differences in prognosis, response to treatment and/or other dependent variables have been demonstrated using appropriate statistical tests [1]. The reliability of these conclusions depends on whether the independent variables being studied, such as diagnostic categories, growth patterns, immunophenotypes, and other features can be assessed in a consistent and reproducible manner. However, multiple studies have demonstrated considerable interobserver variability (IOV) and sometimes even intraobserver variability, that reflect the subjective interpretation of microscopic features and other diagnostic or prognostic variables [2-16]. This problem raises questions about the clinical validity and applicability of conclusions drawn from studies showing prognostic differences among entities that can be diagnosed variably by different pathologists [17-20]. For example, a study evaluating the effect of IOV in the differential diagnosis between usual interstitial pneumonia (UIP) and nonspecific interstitial pneumonia (NSIP) showed that changing diagnostic labels in a manner that simulated IOV in as few as 10% of the cases changed the statistical significance of the prognostic differences estimated in the selected literature [21].

Percent agreement and kappa are the metrics most commonly used in pathology to assess interrater agreement in the interpretation of diagnoses, immunohistochemical results and other test results by two or more observers. Positive and negative percent agreement are the

* Corresponding author at: Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, CA 90048, United States of America.
E-mail address: Alberto.Marchevsky@cshs.org (A.M. Marchevsky).

https://doi.org/10.1016/j.anndiagpath.2020.151561

simplest metrics designed to test for interrater reliability, but they do not consider the prevalence of the entities being rated and/or the possibility that certain agreements can be the result of chance [22]. For example, if the prevalence of tumor A is twice the prevalence of tumor B, some observers are more likely to favor a tumor A diagnosis. Selected investigators have suggested the use of 80% as the minimum interrater agreement level that is considered acceptable for most studies [23].

Jacob Cohen recognized in 1960 that percent agreement is a somewhat unreliable tool by which to assess interrater reliability in psychology and introduced the concept of kappa coefficient to measure the proportion of interrater agreement beyond chance in the interpretation of qualitative, categorical or nominal observations by two observers [24]. However, statisticians such as de Vet et al., have proposed that positive and negative percent agreement are better metrics than Cohen's kappa to compare diagnoses rendered by two raters [22]. Various other coefficients such as Fleiss' kappa (for 3 or more raters), tetrachoric (for dichotomous data and 2 raters), Pearson R, Spearman, Rho, Krippendorff's alpha and other correlation coefficients were proposed to evaluate interrater agreements among 2 or more observers, depending on the particular situation [2,4-6,9,11,13,23,25,26]. Fleiss' kappa, frequently used in the pathology literature, has been applied to IOV studies that compare the observations made by three or more observers [23,25-28].

Landis and Koch proposed a qualitative scale for the interpretation of kappa coefficients in 1977. Their scale has six levels in which kappa coefficients that are < 0 interpreted as no agreement, 0–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–80 as substantial agreement and 0.81–1.0 as almost perfect agreement [9,11,13,18-21,27,29-33]. The use of this scale has been controversial in the psychometric literature and some authors have suggested that conclusions supported by kappa values < 0.67 should be discounted, that conclusions supported by kappa values ranging between 0.67 and 0.80 should be only tentatively accepted and that only conclusions supported by kappa values > 0.80 should be considered as definitive [27].

The scale proposed by Landis and Koch has been widely used in pathology without sufficient discussion as to its applicability for the interpretation of IOV in pathology. For example, is UIP a different clinico-pathologic entity than NSIP that can be diagnosed correctly by some pathologists and incorrectly by others, or are both part of a single clinico-pathologic continuum, in which case each pathologist offers an opinion in a situation where no one is certain about who has issued an accurate diagnosis? How should patients and their clinicians factor this uncertainty into their treatment decisions? Should clinico-pathologic entities that can only be diagnosed with fair, moderate or even substantial agreement be recognized as distinct from each other based on prognostic differences shown in retrospective observational studies or should clinico-pathologic entities only be accepted as distinct if most pathologists can diagnose them consistently and with accuracy similar to that expected for other laboratory tests?

We reviewed the IOV results from a few studies recently published in the lung cancer pathology literature to assess how the authors evaluated kappa values in the conclusions. Metrics commonly used to assess test accuracy in laboratory medicine were also applied to the dataset from one study to evaluate whether the study conclusions were supported when these metrics were applied.

2. Materials and methods

Using a previously described evidence-based approach, the specific questions listed in Table I were formulated to ascertain how pathologists currently assess and interpret IOV in selected problematic areas [34-36]. Four questions were designed to gather an anecdotal impression about the methods being used to assess IOV, the scale being used to interpret the results as clinically relevant, the minimum quantitative levels being used to conclude that IOV would not pose significant clinical problems and the solutions being offered to decrease IOV. Two additional queries were formulated to gain insight into whether pathologists considered themselves or the criteria they were using to define the dependent variables responsible for the IOV in instances where agreement was suboptimal and to investigate whether pathologists suggested that the conclusions previously reported were or might have been biased by the use of dependent variables that could not be identified with acceptable accuracy by different pathologists [37,38]. Answers to these questions were collected from five arbitrarily selected papers that evaluated IOV in a variety of problematic issues in pulmonary cancer pathology [39-43]. These studies investigated reproducibility in each of the following: differential diagnosis of small cell carcinoma, scoring programmed cell death ligand-1 (PDL-1), differential diagnosis between multiple primary lung adenocarcinomas and intrapulmonary metastases, classification of small lung adenocarcinomas into adenocarcinoma in-situ, minimally invasive adenocarcinoma and invasive carcinoma and evaluation of risk of malignancy on cytology specimens.

Table I
Questions on the evaluation and interpretation of interobserver agreement levels in the recent pulmonary pathology literature.

Which method was used to evaluate interobserver agreement levels?
What scale, if any, was used to interpret the results?
Did the study define the minimum quantitative or qualitative level used to conclude that IOV would not pose a significant clinical problem in the interpretation of results?
What solutions were offered to improve agreement levels?
In instances where agreement levels were considered as problematic did the authors conclude that pathologists' perceptions or the definitions of the dependent variables (e.g. classification of tumor into subtypes, characteristics of immunostains, criteria being used to evaluate for stromal invasion) were the root cause of disagreements?
In instances where agreement levels were less than optimal, did the study discuss whether the possibility that conclusions previously reported in the literature were or might be biased by the use of dependent variables that were not well defined?

Metrics commonly applied to assess accuracy in laboratory tests were applied to a dataset from 74 cases from a previous study by Thunnissen et al. [39] using kappa statistics to determine whether the use of immunohistochemistry improved the diagnosis of small cell lung cancer. Briefly, in their study whole slide digital images (WSI) selected from difficult cases had been circulated among 19 pulmonary pathologists from China, Japan, United States, Australia, Argentina, Italy, United Kingdom, and Germany in a manner that closely resembled actual clinical practice [39]. Each participant was provided with 3 opportunities to render “individual diagnostic opinions”, as they studied each tumor stained with hematoxylin and eosin (H&E) and two successive sets of immunostains per case with observer selected or all available immunostains, respectively (“first”, “second” and “third” level “individual diagnostic opinions”). Agreement levels were estimated with kappa statistics and the results interpreted using the Landis and Koch scale [39]. For the current illustration we categorized the “third level” “individual diagnostic opinions” diagnosing cases as small cell carcinoma, large cell neuroendocrine carcinoma, and typical or atypical carcinoid into true and false positive and true and false negative results, using the procedure shown in Table II and the majority consensus of the “third level individual diagnostic opinions” for each case as the gold standard diagnosis for that case. Sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) were calculated using MedCalc software (Ostend, Belgium). “Individual diagnostic opinions” diagnosing cases as non-Hodgkin lymphoma, basaloid squamous cell carcinoma or small round cell sarcoma were excluded because there were fewer than 3 cases in each of these categories.

Table II

Procedure to categorize the diagnostic opinions rendered for each case by multiple observers into true and false positive and true and false negative test results.

Step 1: Create a spreadsheet with 2 columns: individual case number and individual diagnostic opinion and multiple rows including the answers provided by the participants. Sort the data by case number and identify for each case the diagnosis selected by the majority of participants (majority diagnosis).

Step 2: Add a "majority diagnosis" column to the spreadsheet and insert this information into each row as shown in the following example:

Step 3: Add 4 columns to the spreadsheet listing the diagnostic possibilities and classify the majority diagnosis selected by participants as TP, FP, TN or FN, as shown in the following example:

Step 5: Calculate for each of the 4 diagnostic possibilities the percentages of TP, FP, TN and FN and use these data to calculate sensitivity, specificity, positive predictive value and negative predictive value using Medcalc software and standard formulas.

Case number 1	Diagnosis selected by participant	Majority diagnosis
1	Typical carcinoid	Typical carcinoid
1	Atypical carcinoid	Typical carcinoid
1	Typical carcinoid	Typical carcinoid
1	Small cell carcinoma	Typical carcinoid

Case number 1	Diagnosis selected by participant	Majority diagnosis	Typical carcinoid	Atypical carcinoid	Small cell carcinoma	Large cell neuroendocrine carcinoma
1	Typical carcinoid	Typical carcinoid	TP	TN	TN	TN
1	Atypical carcinoid	Typical carcinoid	FN	FP	TN	TN
1	Large cell neuroendocrine carcinoma	Typical carcinoid	TN	TN	TN	FP
1	Small cell carcinoma	Typical carcinoid	TN	TN	FN	TN

3. Results

IOV was evaluated with Fleiss' kappa in 3 of the 5 papers reviewed. One other paper used Cohen's kappa while the other did not specify the statistical method selected to estimate kappa scores. Three of the 5 papers interpreted the results using the Landis-Koch scale and one modified the scale classifying kappa scores of 0.61–0.80 as good, as initially reported by Cohen, rather than as "substantial". The fifth paper used a modification of the scale proposed by Cohen. None of the five papers explicitly indicated or proposed a minimum kappa score that would reasonably exclude the possibility that IOV could bias results in a clinically significant manner. One of the papers considered kappa scores < 0.40 as outliers while the others did not propose a particular kappa score to determine whether the level of IOV would be acceptable in clinical practice. The root cause of suboptimal IOV was attributed to overlapping diagnostic features that required the development of better criteria in all 5 studies. Two of the studies also opined that interpretation of current criteria by pathologists was part of the problem. Solutions proposed to decrease IOV included the use of immunohistochemistry, molecular studies and machine learning and development of more explicit microscopic features to distinguish the categories under investigation. None of the papers considered the possibility that the problem did not reside with the diagnostic criteria or pathologists but with the fact that the entities being differentiated are not entirely different from each other and that IOV could be resolved by combining them into fewer categories that could be recognized with greater accuracy.

Table III shows the percent agreement between "individual diagnostic opinions" and majority consensus diagnoses calculated from the data reported by Thunnissen et al. [20]. The table also shows the kappa scores reported in that study. Table IV shows the results of sensitivity, specificity, and positive and negative predictive value calculations using the majority consensus diagnoses as ground truth. Although the kappa scores in Table III are "Fair" to "Good" kappa scores, all four neuroendocrine neoplasms were diagnosed with $> 90\%$ specificity by the study participants (Table IV).

4. Discussion

While our review of only five studies from the pathology literature certainly does not represent a comprehensive evaluation of the methodologies being used to evaluate IOV and interpret the clinical implications, the results are consistent with our anecdotal impression about this literature. Most studies have evaluated IOV with kappa statistics and although they generally report % agreement, they use kappa scores interpreted with the Landis-Koch scale or minor modifications of this scale. However, there appears to be no consensus about how to interpret kappa scores in the context of clinical practice. None of the studies we sampled cited the existence of an expert opinion or evidence-based rule that could be used to determine which or even whether a particular level of kappa score, such as substantial, moderate or good would indicate that pathologists can establish particular differential diagnosis with an acceptable level of accuracy to ensure that patients are diagnosed in a reproducible manner. Interestingly, based on our analysis of the data from the study by Thunnissen et al., it would appear there is no need for such a rule for the diagnosis of lung neuroendocrine neoplasms. Indeed, diagnostic specificity was $> 90\%$ for the four neoplasms, in spite of kappa scores that were quite variable, suggesting that kappa statistics are of little clinical value when comparing the opinion of multiple raters against a "golden reference" diagnosis. We elected not to evaluate how best to compare diagnostic opinions among two raters by calculating the positive and negative percent agreement for each pairwise comparison of the raters, as estimates of median, minimum and maximum agreements for each diagnosis are not appropriate to compare the different raters to a gold reference diagnosis. The problem of how best to compare diagnosis by two raters was discussed in 2013

Table III
Fleisher kappa values and % agreement levels with majority consensus diagnoses.

Diagnosis (n = number of observations)	Kappa score*	Interpretation of kappa score**	% Agreement***
Small cell lung carcinoma (n = 589)	0.60	Moderate	68.9
Typical carcinoid (n = 342)	0.74	Good	84.5
Large cell neuroendocrine carcinoma (n = 247)	0.49	Moderate	73.3
Atypical carcinoid (n = 133)	0.30	Fair	54.9

* From Thunnissen et al. [39].
** Kappa values were interpreted using the standard Landis and Koch [29].
*** Majority diagnoses were used as "gold standard" in this calculation.

Table IV
Metrics to estimate the accuracy of diagnoses using majority consensus diagnoses as the gold standard for "true positives".

Diagnosis	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Small cell lung carcinoma	68.9%	92.8%	87.3%	80.6%
Typical carcinoid	84.5%	96.7%	89.2%	95.1%
Large cell neuroendocrine carcinoma	73.3%	92.9%	68.8%	94.2%
Atypical carcinoid	54.9%	93.0%	45.1%	95.2%

by de Vet et al. [22]. The study concluded that clinicians are justified in their unhappiness about using Cohen's kappa for comparisons of diagnoses among two raters and favored the use of positive and negative percent agreement for such estimates.

Analysis of the data collected for the study of Thunnissen et al. highlights the problems related to the identification of a "gold reference" to use as true and false positive and negative results when calculating sensitivity, specificity, NPV and PPV. Thunnissen et al. did not record the diagnoses rendered by the pathologists who provided the cases for investigation. As the "gold reference" or "ground truth" for each case we elected to use the majority consensus diagnoses among the study participants, although this method does not offer assurance that the majority was correct. Indeed, while almost all participants agreed on the diagnosis in certain cases, other cases were more problematic as evidenced by only a slight majority concurring with the "gold reference". Future studies comparing the IOV between multiple observers and the diagnoses rendered by the pathologists who submitted the cases or by experts could provide a better study design to evaluate IOV among multiple raters against an accurate "gold reference" diagnoses.

Recent studies published in 2020 also illustrate the lack of guidelines about how to evaluate the accuracy of immunohistochemistry interpretations by different pathologists. For example, Thunnissen et al. and Huang et al. compared agreement rates among pathologists evaluating programmed death-ligand 1 expression by immunohistochemistry and ROS-1 by fluorescence in situ hybridization using positive and negative % agreement, while Williams et al. elected to analyze agreement in scores using intraclass correlation coefficients and concordance in patient's classification using Fleiss' kappa [40,44,45]. Although the application of different statistical methods is correct, this variability complicates the comparison of results across different studies comparing similar problems

There has been limited discussion about the effect of variability and uncertainty in diagnostic classifications, prognostic models, evaluation of the results of molecular studies and clinical trials. As recently reviewed by McHugh and other authors from the United States Food and Drug Administration (FDA), misclassifications by as little as 5% can be sufficient to significantly invalidate estimates of specificity, sensitivity and area under receiver operating curves [46]. Other studies have shown the effect of misclassifications on prognosis and other health related prediction models, but to our knowledge there are no evidence or expert opinion guidelines on how to control for this problem in future studies proposing new pathologic entities based on prognostic differences or evaluating the utility of new therapeutic options, by diagnosis [47,48].

In summary, our review of literature shows that kappa statistics have limited clinical applicability in pathology and suggests the need for guidelines that would help standardize the evaluation of IOV in a manner that would facilitate comparison of different studies and performance of meta-analysis. We concur with de Vet el al's conclusions that positive and negative agreement levels are the preferred metrics for evaluation of IOV among two raters and propose that estimates of diagnostic specificity, sensitivity and positive and negative predictive values against a "ground truth" are most useful to evaluate IOV among multiple raters.

Acknowledgements

The authors thank Prof. Dr. I. Wistuba and Prof. Dr. K. Kerr, respective chairs of the IASLC pathology panel for their permission to use the data of the 'Small Cell Carcinoma Lung Cancer-study': "The Use of Immunohistochemistry Improves the Diagnosis of Small Cell Lung Cancer and Its Differential Diagnosis. An International Reproducibility Study in a Demanding Set of Cases" [39].

References

[1] Khan KS, Chien PF. Evaluation of a clinical test. I: assessment of reliability. BJOG 2001;108:562–7.
[2] Thompson LDR, Poller DN, Kakudo K, Burchette R, Nikiforov YE, Seethala RR. An international interobserver variability reporting of the nuclear scoring criteria to diagnose noninvasive follicular thyroid neoplasm with papillary-like nuclear features: a validation study. Endocr Pathol 2018;29:242–9.
[3] Osmond A, Li-Chang H, Kirsch R, Divaris D, Falck V, Liu DF, et al. Interobserver variability in assessing dysplasia and architecture in colorectal adenomas: a multicentre Canadian study. J Clin Pathol 2014;67:781–6.
[4] Hoffman A, Rey JW, Mueller L, Hansen T, Goetz M, Tresch A, et al. Analysis of interobserver variability for endomicroscopy of the gastrointestinal tract. Dig Liver Dis 2014;46:140–5.
[5] Chebib I, Rao RA, Wilbur DC, Tambouret RH. Using the ASC:SIL ratio, human papillomavirus, and interobserver variability to assess and monitor cytopathology fellow training performance. Cancer Cytopathol 2013;121:638–43.
[6] van den Einden LC, de Hullu JA, Massuger LF, Grefte JM, Bult P, Wiersma A, van Engen-van Grunsven AC, Sturm B, Bosch SL, Hollema H, Bulten J. Interobserver variability and the effect of education in the histopathological diagnosis of differentiated vulvar intraepithelial neoplasia. Mod Pathol 2013; 26, 874–880.
[7] Eriksson H, Frohm-Nilsson M, Hedblad MA, Hellborg H, Kanter-Lewensohn L, Krawiec K, et al. Interobserver variability of histopathological prognostic parameters in cutaneous malignant melanoma: impact on patient management. Acta Derm Venereol 2013;93:411–6.
[8] Wolfson WL. Interobserver variability among expert uropathologists. Am J Surg Pathol 2009;33:801. [author reply 801-802].
[9] Evans AJ, Henry PC, Van der Kwast TH, Tkachuk DC, Watson K, Lockwood GA, et al. Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens. Am

J Surg Pathol 2008;32:1503–12.
[10] Montgomery E. Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? Arch Pathol Lab Med 2005;129:174–6.
[11] Verkooijen HM, Peterse JL, Schipper ME, Buskens E, Hendriks JH, Pijnappel RM, et al. Interobserver variability between general and expert pathologists during the histopathological assessment of large-core needle and open biopsies of non-palpable breast lesions. Eur J Cancer 2003;39:2187–91.
[12] Chhieng DC, Talley LI, Roberson J, Gatscha RM, Jhala NC, Elgert PA. Interobserver variability: comparison between liquid-based and conventional preparations in gynecologic cytology. Cancer 2002;96:67–73.
[13] Odze RD, Goldblum J, Noffsinger A, Alsaigh N, Rybicki LA, Fogt F. Interobserver variability in the diagnosis of ulcerative colitis-associated dysplasia by telepathology. Mod Pathol 2002;15:379–86.
[14] Cramer SF. Interobserver variability in dermatopathology. Arch Dermatol 1997;133:1033–6.
[15] Raab SS, Robinson RA, Snider TE, McDaniel HL, Sigman JD, Leigh CJ, et al. Telepathologic review: utility, diagnostic accuracy, and interobserver variability on a difficult case consultation service. Mod Pathol 1997;10:630–5.
[16] Sheibani K, Nathwani BN, Swartz WG, Ben-Ezra J, Brownell MD, Burke JS, et al. Variability in interpretation of immunohistologic findings in lymphoproliferative disorders by hematopathologists. A comprehensive statistical analysis of interobserver performance. Cancer 1988;62:657–64.
[17] Thunnissen E, Witte BI, Nicholson AG. all a. Reproducibility of histopathological diagnosis in poorly differentiated NSCLC: an international multiobserver study. J Thorac Oncol 2015;10:e4.
[18] Thunnissen E, Noguchi M, Aisner S, Beasley MB, Brambilla E, Chirieac LR, et al. Reproducibility of histopathological diagnosis in poorly differentiated NSCLC: an international multiobserver study. J Thorac Oncol 2014;9:1354–62.
[19] Thunnissen E, Boers E, Heideman DA, Grunberg K, Kuik DJ, Noorduin A, et al. Correlation of immunohistochemical staining p63 and TTF-1 with EGFR and K-ras mutational spectrum and diagnostic reproducibility in non small cell lung carcinoma. Virchows Arch 2012;461:629–38.
[20] Thunnissen E, Beasley MB, Borczuk AC, Brambilla E, Chirieac LR, Dacic S, et al. Reproducibility of histopathological subtypes and invasion in pulmonary adenocarcinoma. An international interobserver study. Mod Pathol 2012;25:1574–83.
[21] Marchevsky AM, Gupta R. Interobserver diagnostic variability at “moderate” agreement levels could significantly change the prognostic estimates of clinicopathologic studies: evaluation of the problem using evidence from patients with diffuse lung disease. Ann Diagn Pathol 2010;14:88–93.
[22] de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen’s kappa. BMJ 2013;346:f2125.
[23] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276–82.
[24] A JC. Coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:36–7.
[25] Banerjee MC, M; McSweeney, L; Sinha, D. Beyond kappa: a review of interrater agreement measures. Canadian J Statistics 1999; 27, 3–23.
[26] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.
[27] Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012;8:23–34.
[28] Fleiss JL, Spitzer RL, Endicott J, Cohen J. Quantification of agreement in multiple psychiatric diagnosis. Arch Gen Psychiatry 1972;26:168–71.
[29] Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363–74.
[30] Wright KC, Melia J, Moss S, Berney DM, Coleman D, Harnden P. Measuring interobserver variation in a pathology EQA scheme using weighted kappa for multiple readers. J Clin Pathol 2011;64:1128–31.
[31] Venkataraman G, Ananthanarayanan V, Paner GP. Accessible calculation of multirater kappa statistics for pathologists. Virchows Arch 2006;449:272.
[32] Thomson TA, Hayes MM, Spinelli JJ, Hilland E, Sawrenko C, Phillips D, et al. HER-2/neu in breast cancer: interobserver variability and performance of immunohistochemistry with 4 antibodies compared with fluorescent in situ hybridization. Mod Pathol 2001;14:1079–86.
[33] Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS 1989;97:689–98.
[34] Marchevsky AM, Walts AE, Bose S, Gupta R, Fan X, Frishberg D, et al. Evidence-based evaluation of the risks of malignancy predicted by thyroid fine-needle aspiration biopsies. Diagn Cytopathol 2010;38:252–9.
[35] Herbst J, Jenders R, McKenna R, Marchevsky A. Evidence-based criteria to help distinguish metastatic breast cancer from primary lung adenocarcinoma on thoracic frozen section. Am J Clin Pathol 2009;131:122–8.
[36] Gupta R, Marchevsky AM, McKenna RJ, Wick M, Moran C, Zakowski MF, et al. Evidence-based pathology and the pathologic evaluation of thymomas: transcapsular invasion is not a significant prognostic feature. Arch Pathol Lab Med 2008;132:926–30.
[37] Marchevsky AM. Evidence-based medicine in pathology: an introduction. Semin Diagn Pathol 2005;22:105–15.
[38] Marchevsky AM, Wick MR. Evidence-based medicine, medical decision analysis, and pathology. Hum Pathol 2004;35:1179–88.
[39] Thunnissen E, Borczuk AC, Flieder DB, Witte B, Beasley MB, Chung JH, et al. The use of immunohistochemistry improves the diagnosis of small cell lung cancer and its differential diagnosis. An international reproducibility study in a demanding set of cases. J Thorac Oncol 2017;12:334–46.
[40] Williams GH, Nicholson AG, Snead DRJ, Thunnissen E, Lantuejoul S, Cane P, et al. Interobserver reliability of programmed cell death Ligand-1 scoring using the VENTANA PD-L1 (SP263) assay in NSCLC. J Thorac Oncol 2020;15:550–5.
[41] Nicholson AG, Torkko K, Viola P, Duhig E, Geisinger K, Borczuk AC, et al. Interobserver variation among pathologists and refinement of criteria in distinguishing separate primary tumors from intrapulmonary metastases in lung. J Thorac Oncol 2018;13:205–17.
[42] Shih AR, Uruga H, Bozkurtlar E, Chung JH, Hariri LP, Minami Y, et al. Problems in the reproducibility of classification of small lung adenocarcinoma: an international interobserver study. Histopathology 2019;75:649–59.
[43] Hiroshima K, Yoshizawa A, Takenaka A, Haba R, Kawahara K, Minami Y, et al. Cytology reporting system for lung cancer from the Japan Lung Cancer Society and Japanese Society of Clinical Cytology: an interobserver reproducibility study and risk of malignancy evaluation on cytology specimens. Acta Cytol 2020:1–11.
[44] Thunnissen E, Kerr KM, Dafni U, Bubendorf L, Finn SP, Soltermann A, et al. Programmed death-ligand 1 expression influenced by tissue sample size. Scoring based on tissue microarrays’ and cross-validation with resections, in patients with, stage I-III, non-small cell lung carcinoma of the European Thoracic Oncology Platform Lungscape cohort. Mod Pathol 2020;33:792–801.
[45] Huang RSP, Smith D, Le CH, Liu WW, Ordinario E, Manohar C, et al. Correlation of ROS1 immunohistochemistry with ROS1 fusion status determined by fluorescence in situ hybridization. Arch Pathol Lab Med 2020;144:735–41.
[46] McHugh LC, Snyder K, Yager TD. The effect of uncertainty in patient classification on diagnostic performance estimations. PLoS One 2019;14:e0217146.
[47] Sposto R, Preston DL, Shimizu Y, Mabuchi K. The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in A-bomb survivors. Biometrics 1992;48:605–17.
[48] van Doorn S, Brakenhoff TB, Moons KGM, Rutten FH, Hoes AW, Groenwold RHH, et al. The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: a case study of the CHA2DS2-VASc score in atrial fibrillation. Diagn Progn Res 2017;1:18.