ORIGINAL ARTICLE

Quantifying Interrater Agreement and Reliability Between Thoracic Pathologists: Paradoxical Behavior of Cohen's Kappa in the Presence of a High Prevalence of the Histopathologic Feature in Lung Cancer

Kay See Tan, PhD,[a,*] Yi-Chen Yeh, MD,[b] Prasad S. Adusumilli, MD, FACS,[c] William D. Travis, MD[d]

[a]*Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York*
[b]*Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei, Taiwan*
[c]*Thoracic Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, New York*
[d]*Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, New York*

Received 16 October 2023; revised 28 November 2023; accepted 10 December 2023
Available online - 16 December 2023

ABSTRACT

Introduction: Cohen's kappa is often used to quantify the agreement between two pathologists. Nevertheless, a high prevalence of the feature of interest can lead to seemingly paradoxical results, such as low Cohen's kappa values despite high "observed agreement." Here, we investigate Cohen's kappa using data from histologic subtyping assessment of lung adenocarcinomas and introduce alternative measures that can overcome this "kappa paradox."

Methods: A total of 50 frozen sections from stage I lung adenocarcinomas less than or equal to 3 cm in size were independently reviewed by two pathologists to determine the absence or presence of five histologic patterns (lepidic, papillary, acinar, micropapillary, solid). For each pattern, observed agreement (proportion of cases with concordant "absent" or "present" ratings) and Cohen's kappa were calculated, along with Gwet's AC1.

Results: The prevalence of any amount of the histologic patterns ranged from 42% (solid) to 97% (acinar). On the basis of Cohen's kappa, there was substantial agreement for four of the five patterns (lepidic, 0.65; papillary, 0.67; micropapillary, 0.64; solid, 0.61). Acinar had the lowest Cohen's kappa (0.43, moderate agreement), despite having the highest observed agreement (88%). In contrast, Gwet's AC1 values were close to or higher than Cohen's kappa across patterns (lepidic, 0.64; papillary, 0.69; micropapillary, 0.71; solid, 0.73; acinar, 0.85). The proportion of positive versus negative agreement was 93% versus 50% for acinar.

Conclusions: Given the dependence of Cohen's kappa on feature prevalence, interrater agreement studies should include complementary indices such as Gwet's AC1 and proportions of specific agreement, especially in settings with a high prevalence of the feature of interest.

Keywords: Interobserver coefficient; Reproducibility; Predominant histologic subtypes; Diagnostic accuracy; Performance metrics; Sensitivity and specificity

*Corresponding author.

Disclosure: Dr. Adusumilli's laboratory receives research support from ATARA Biotherapeutics. The remaining authors declare no conflict of interest.

Address for correspondence: Kay See Tan, PhD, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065. E-mail: tank@mskcc.org

Cite this article as: Tan KS, Yeh YC, Adusumilli PS, Travis WD. Quantifying interrater agreement and reliability between thoracic pathologists: paradoxical behavior of Cohen's kappa in the presence of a high prevalence of the histopathologic feature in lung cancer. *JTO Clin Res Rep*. 2024;5:100618.

ISSN: 2666-3643

https://doi.org/10.1016/j.jtocrr.2023.100618

Introduction

Interrater agreement and reliability are key metrics to determine the reproducibility of diagnoses, immunohistochemical results, and other test results such as molecular assays in surgical pathology. If two pathologists can reliably apply a criterion or tool to make the same assessment on the same specimen, the interrater agreement will be high and can serve as evidence of reliable ratings. If the ratings are highly discordant, then either the tool is not useful or the raters require additional training. The statistical measure most widely used to quantify the agreement between pathologists is Cohen's kappa.[1] Cohen's kappa reflects the agreement beyond that which occurs by chance (i.e., chance corrected). Despite its popularity, Cohen's kappa has been found to produce paradoxical results under certain circumstances.[2,3] Paradoxical results occur when a high level of agreement is accompanied by a low kappa value, leading to seemingly counterintuitive conclusions. In the present study, we review Cohen's kappa statistic and assess its limitations using data from a published study of surgical pathology in lung cancer. We provide practical recommendations and propose alternative measures of agreement for future studies of interrater agreement.

Materials and Methods

Patient Data and Study Design

The present study uses data from a surgical pathology study by Yeh et al.[4] that focused on stage I lung adenocarcinomas less than or equal to 3 cm in size. Details regarding patient selection, study methods, and evaluation of surgical specimens are reported in the previous study.[4] Data were collected under a protocol (IRB 17-630) approved by the Institutional Review Board at Memorial Sloan Kettering Cancer Center, which included a waiver of informed consent.

In brief, patients with lung adenocarcinoma less than or equal to 3 cm in size who underwent surgical resection from 1995 to 2009 were identified from the prospectively curated Memorial Sloan Kettering Cancer Center Thoracic Service database. Original permanent and frozen section slides were available for a cohort of 361 patients. By analyzing various subsets of the 361-patient cohort, Yeh et al.[4] investigated the strengths and limitations of frozen sections for the accurate identification of prognostically important histologic features. In particular, a subset of 50 patients was randomly selected from the full cohort of 361 patients and independently reviewed by three pathologists to determine the presence or absence of lepidic, acinar, papillary, micropapillary, and solid patterns on frozen sections.[4,5]

The present study uses data from this same set of 50 frozen sections. For the purpose of illustration, we use the ratings from two (instead of three) pathologists. On the basis of these ratings, various agreement measures are presented, which are as follows: "observed agreement" (the proportion of cases with the same ratings from both raters), "chance agreement" (the probability of two raters agreeing by random chance), and "chance-corrected agreement" (agreement metrics that adjust for chance agreement, such as Cohen's kappa and Gwet's AC1).

Cohen's Kappa

The equation for Cohen's kappa is presented in Figure 1. Cohen's kappa ranges from 0 to 1, where higher values indicate greater interrater agreement. The degree of agreement is conventionally categorized as poor (kappa $\leq$ 0.20), fair (0.21 $\leq$ kappa $\leq$ 0.40), moderate (0.41 $\leq$ kappa $\leq$ 0.60), substantial (0.61 $\leq$ kappa $\leq$ 0.80), and almost perfect (0.81 $\leq$ kappa $\leq$ 1.00).[6]

Gwet's AC1

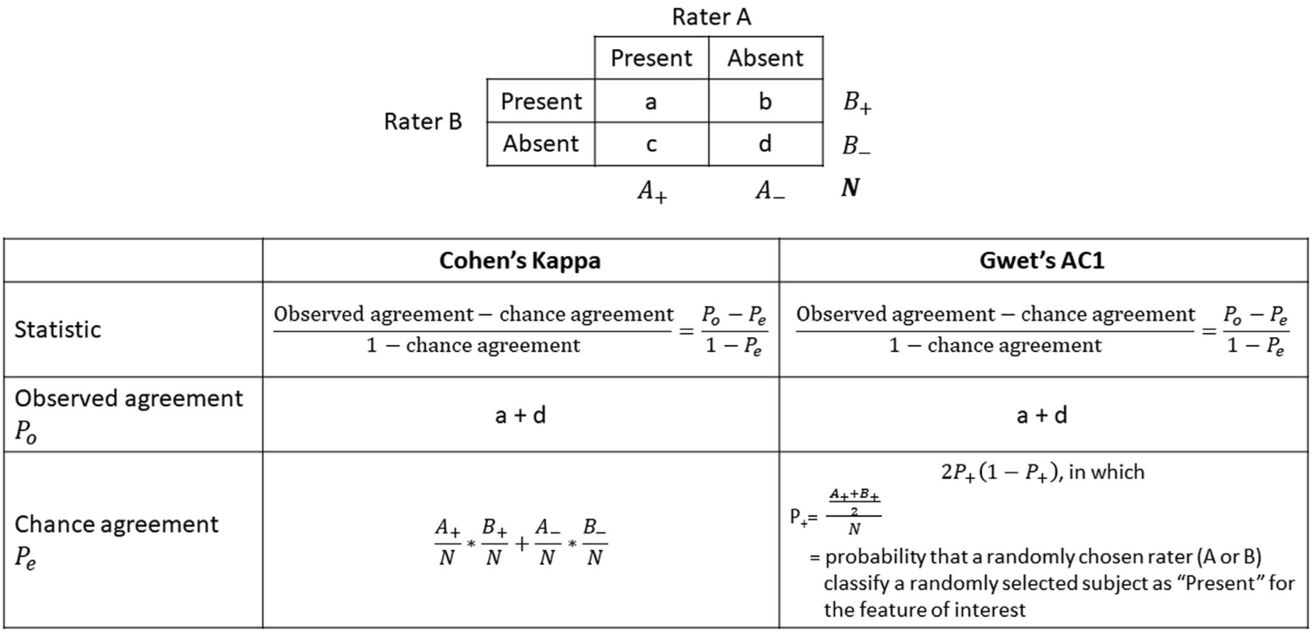
Gwet's AC1 is calculated using the formula presented in Figure 1. Similar to Cohen's kappa, Gwet's AC1 attempts to remove the chance agreement from the observed agreement, using the same structure of (observed agreement – chance agreement) / (1 – chance agreement).

Positive and Negative Agreement

The proportion of specific agreement includes two separate indices, P_{pos} (positive agreement) and P_{neg} (negative agreement). P_{pos} refers to the proportion of cases that were classified as positive (i.e., the feature of interest is present) among the average number of positive ratings between the two pathologists, whereas P_{neg} refers to the average proportional negative agreement. In accordance with the notations in Figure 1, the number of positive readings is A_+ for rater A and B_+ for rater B. Hence, positive agreement is calculated as $P_{pos} = a / [(A_+ + B_+) / 2]$, and negative agreement is calculated as $P_{neg} = d / [(A_- + B_-) / 2]$.

Statistical Analysis

Patient characteristics are summarized as frequency and percentage for categorical variables and as median (25th–75th percentiles) for continuous variables. On the basis of the "absent" or "present" ratings across the 50 frozen sections for each histologic pattern, we calculated the observed agreement between the two pathologists, Cohen's kappa, and Gwet's AC1. We also derived the observed proportion of positive and negative agreement ($P_{pos} : P_{neg}$). In addition, we determined the prevalence of each pattern (prevalence of the feature of interest) on the basis of the proportion of cases with the feature



		Rater A		
		Present	Absent	
Rater B	Present	a	b	B_+
	Absent	c	d	B_-
		A_+	A_-	N

	Cohen's Kappa	Gwet's AC1
Statistic	$\frac{\text{Observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}} = \frac{P_o - P_e}{1 - P_e}$	$\frac{\text{Observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}} = \frac{P_o - P_e}{1 - P_e}$
Observed agreement P_o	a + d	a + d
Chance agreement P_e	$\frac{A_+}{N} * \frac{B_+}{N} + \frac{A_-}{N} * \frac{B_-}{N}$	$2P_+(1 - P_+)$, in which $P_+ = \frac{\frac{A_+ + B_+}{2}}{N}$ = probability that a randomly chosen rater (A or B) classify a randomly selected subject as "Present" for the feature of interest

Figure 1. Calculation of chance-corrected agreement (Cohen's kappa and Gwet's AC1 statistics) on the basis of observed and chance agreement.

present in the full cohort.[4] Observed agreement, Cohen's kappa, and Gwet's AC1 were calculated using the *immer*[7,8] and *epiR*[9] packages in R (version 4.1.2, R Corporation, Vienna, Austria). For comparison, we present two additional alternative agreement metrics, which are as follows: Aickin's α[10] and B statistic[11] from the *immer*[7,8] and *vcd*[12] packages in R.

Results

Clinicopathologic Characteristics of the Patients

The characteristics of the 50 included patients are summarized in Table 1. On the basis of the full cohort of 361 patients previously reported by Yeh et al.,[4] the prevalence of each pattern (prevalence of the feature of interest) ranged from 42% for solid pattern to 97% for acinar pattern (Table 2).

Interrater Agreement for the Presence of Histologic Patterns Using Frozen Sections

The observed agreement between the two pathologists was high across all five histologic patterns, ranging from 82% for lepidic pattern to 84% for acinar pattern (Table 2).

The conventional approach (i.e., using Cohen's kappa) indicated substantial agreement for four of the five histologic patterns (kappa: lepidic, 0.65; papillary, 0.67; micropapillary, 0.64; solid, 0.61). The lowest Cohen's kappa was 0.43, for acinar pattern, which corresponds to moderate agreement.

Gwet's AC1 values were close to or higher than Cohen's kappa across all five patterns (Gwet's AC1: lepidic, 0.64; papillary, 0.69; micropapillary, 0.71; solid, 0.73). In particular, Gwet's AC1 for acinar pattern (0.85) was the highest across all five patterns. Although not the focus of the current study, B statistics and Aickin's α were similar to Gwet's AC1 except for acinar and micropapillary, in which Aickin's α values were in between Cohen's kappa and Gwet's AC1.

Influence of the Prevalence of the Feature of Interest on Interrater Agreement

The prevalence of each histologic pattern is presented in Table 2. For the four patterns with a Cohen's kappa greater than 0.6 (lepidic, papillary, micropapillary, and solid), the prevalence of any amount of each pattern was between 42% and 75%; for the pattern with the lowest Cohen's kappa (acinar; Cohen's kappa, 0.43), the prevalence was 97%.

Table 1. Patient Characteristics (N = 50)

Characteristics	Median (25th-75th Percentile) or n (%)
Age at surgery, y	66 (60-73)
Sex	
Female	28 (56)
Male	22 (44)
Pathologic stage	
1A	49 (98)
1B	1 (2.0)

Table 2. Interobserver Agreement Between Two Pathologists for the Presence or Absence of Histologic Patterns Using Frozen Sections

Features	Ratings by Two Pathologists			Observed Agreement, %	Cohen's Kappa	Prevalence of Feature,[a] %	P_{pos} (95% CI)[b]	P_{neg} (95% CI)[b]	Gwet's AC1	B Statistic	Aickin's α
Lepidic		Present	Absent	82	0.65	69	80 (64-90)	84 (71-92)	0.64	0.70	0.78
	Present	18	9								
	Absent	0	23								
Acinar		Present	Absent	88	0.43	97	93 (86-97)	50 (17-78)	0.85	0.86	0.66
	Present	41	2								
	Absent	4	3								
Papillary		Present	Absent	84	0.67	75	80 (63-90)	87 (75-94)	0.69	0.73	0.73
	Present	16	7								
	Absent	1	26								
Micropapillary		Present	Absent	84	0.64	47	76 (57-89)	88 (77-94)	0.71	0.73	0.67
	Present	13	4								
	Absent	4	29								
Solid		Present	Absent	84	0.61	42	71 (49-86)	89 (79-95)	0.73	0.76	0.72
	Present	10	7								
	Absent	1	32								

CI, confidence interval; P_{neg}, negative agreement; P_{pos}, positive agreement.

[a] The prevalence of the feature was derived from the full cohort of 361 patients from the study from Yeh et al.[4]

[b] The 95% CIs around P_{pos} and P_{neg} reflect Beta (1,1) prior.

The details for solid pattern versus acinar pattern illustrate the influence of the prevalence of the feature of interest on interrater agreement. The prevalence of solid pattern was 42%, and the observed agreement between the two pathologists was 84%. Cohen's kappa resulted in a chance-corrected agreement of 0.61, similar to Gwet's AC1 of 0.73. In contrast, the prevalence of acinar pattern was 97%. Even with a high observed agreement of 88% between the two pathologists, Cohen's kappa was 0.43, compared with Gwet's AC1 of 0.85 (which was closer to the observed agreement).

Distinguishing Between Positive and Negative Agreement

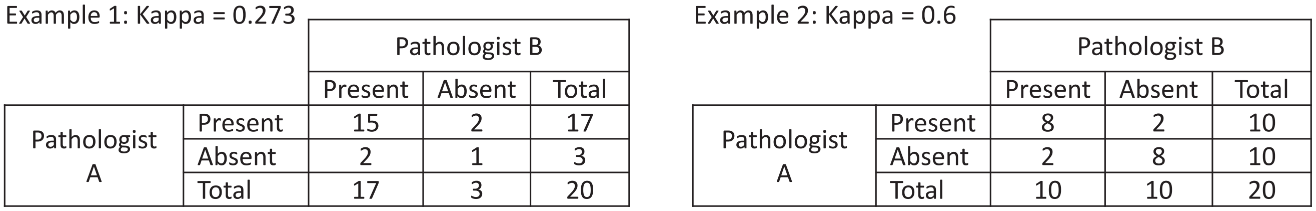
When lepidic pattern was assessed, the average number of "present" and "absent" ratings was 22.5 and 27.5 of 50 cases, respectively. Hence, the proportion of "present" ratings that were concordant between the two pathologists (P_{pos}) was 80% (18 of 22.5), and the proportion of "absent" ratings that were concordant (P_{neg}) was 84% (23 of 27.5). High P_{pos} and P_{neg} values were similarly observed for papillary, micropapillary, and solid patterns.

In contrast, when acinar pattern was assessed, the average number of "present" and "absent" ratings was 44 and six of 50 cases, respectively, reflecting a high prevalence of the pattern. The six discordant ratings resulted in P_{pos} of 93% (41 of 44) and P_{neg} of 50% (three of six). This implies that, in practice, if one pathologist rates the case as "absent" for acinar pattern, it may be worthwhile to obtain the opinion of a second pathologist. In the case of a "present" rating, however, the probability that the second pathologist agrees is 93%.

Discussion

Cohen's kappa is routinely used to determine interrater agreement between two raters. The primary idea underlying Cohen's kappa is that part of the observed agreement between two raters is attributable to chance—that is, that the two raters agree (whether the feature of interest is present or absent) simply because of chance. Cohen's kappa adjusts for this chance agreement to derive a chance-corrected agreement. Two examples using data from the study population are provided subsequently to illustrate the potential limitations of Cohen's kappa.

In example 1, pathologist A and pathologist B agree with each other on 16 of 20 frozen section slides. On 15 of the 16 slides, both pathologists observed the feature of interest, and on one slide, both pathologists did not observe the feature of interest (Fig. 2; example 1). Therefore, the observed agreement is as follows: (15 + 1) / 20 = 0.8. Nevertheless, pathologist A may have



Example 1: Kappa = 0.273

		Pathologist B		
		Present	Absent	Total
Pathologist A	Present	15	2	17
	Absent	2	1	3
	Total	17	3	20

Example 2: Kappa = 0.6

		Pathologist B		
		Present	Absent	Total
Pathologist A	Present	8	2	10
	Absent	2	8	10
	Total	10	10	20

Figure 2. Summary of ratings by two pathologists; both examples have 80% observed agreement between the two raters.

agreed with pathologist B simply by chance even if neither pathologist had scrutinized the frozen sections. To calculate the chance agreement, note that pathologist A found that 17 of 20 slides had the feature present and three of 20 slides had the feature absent. Thus, pathologist A said “present” 85% of the time and pathologist B said “present” 85% of the time. Consequently, the probability that both pathologists said “present” was $0.85 \times 0.85 = 0.7225$, and the probability that both pathologists said “absent” was $0.15 \times 0.15 = 0.0225$. The overall chance agreement is, therefore, $0.7225 + 0.0225 = 0.745$, meaning that 74.5% of agreement between the pathologists is attributable to chance. Following the formula in Figure 1, Cohen's kappa is calculated as (observed agreement – chance agreement) / (1 – chance agreement), which yields $\kappa = (0.8 - 0.745) / (1 - 0.745) \approx 0.22$; this is considered poor to fair.

In example 2 (Fig. 2), the observed agreement is exactly the same as in example 1—80% (16 of 20 cases in agreement)—but Cohen's kappa is much higher because of a smaller chance agreement. The chance agreement in example 2 is $(0.5 \times 0.5) + (0.5 \times 0.5) = 0.5$, which yields a Cohen's kappa as follows: $\kappa = (0.8 - 0.5) / (1 - 0.5) = 0.6$; this is considered moderate agreement. Despite that both examples were derived from tables with an observed agreement of 80%, example 1 had a lower kappa value (kappa = 0.27 in example 1 versus kappa = 0.6 in example 2). This discrepancy is because of a markedly different prevalence of the feature of interest (17 / 20 = 85% in example 1 versus 10 / 20 = 50% in example 2). When the prevalence of the feature of interest is close to 50%, as in example 2, the resulting kappa value is closer to the observed agreement. In contrast, when the prevalence of the feature of interest is either very high (close to 100%) or very low (close to 0%), as in example 1, the kappa value seems to be counterintuitively low.

To avoid the paradoxical results that can occur with Cohen's kappa under certain circumstances, Gwet[13] proposed a new agreement measure called the “first-order agreement coefficient” or AC1. The primary difference between Cohen's kappa and Gwet's AC1 lies in the calculation of chance agreement, which is based on the chance that raters may agree on a rating despite the fact that one or both of them may have made a random classification. Random ratings can occur when the rater is uncertain about how to classify a specimen (perhaps when the specimen's characteristics do not match the rating instructions) and hence randomly assigns “present” for the feature of interest. In a situation where Cohen's kappa is low despite a high level of overall agreement, Gwet's AC1 has been introduced as a “paradox-resistant” alternative to Cohen's kappa.[14]

As revealed in the present study, Gwet's AC1 provides a chance-corrected agreement coefficient that is more in line with observed agreement, compared with Cohen's kappa. Despite its popularity, Cohen's kappa has its drawbacks, particularly in the setting of a high prevalence of the feature of interest. Cohen's kappa assumes that agreement is at random and, hence, captures the agreement beyond that occurring at random. Conversely, Gwet's AC1 acknowledges that agreement between observers is not totally at random—that is, there will be cases where the feature is truly present that will be easy to reach agreement on, there will be cases where the feature is truly absent that will be easy to reach agreement on, and there will be cases for which it will be difficult to reach agreement. Taking this perspective into consideration, Gwet's AC1 avoids the overpenalization that results with Cohen's kappa simply as a consequence of a high prevalence of the feature of interest.

The findings in this illustrative study, by the use of the ratings of two pathologists, can be extended to the setting of multiple response categories and multiple raters. Instead of only two possible attributions (“present” or “absent”), the response categories can be ordinal, such as “absent,” “low,” “intermediate,” and “high.” Cohen's kappa has been extended to handle such settings using the weighted kappa.[15] Furthermore, whereas Cohen's kappa applies only to two raters, Light's kappa[16] can be applied in the setting of multiple raters. Both Cohen's kappa and Light's kappa assume that a fixed number of raters are rating identical cases; in contrast, Fleiss' kappa[17] is a more flexible approach that can be applied to any number of raters rating different cases. Similarly, Gwet's AC1 has also been extended to accommodate multiple raters.[18]

To the best of our knowledge, Gwet's AC1 has never been compared with Cohen's kappa in the context of lung cancer pathology. Nevertheless, discussions surrounding the paradox of low Cohen's kappa despite high observed agreement have been ongoing. Whereas some have cautioned against the use of Cohen's kappa in these settings,[2,3] others have argued for continued support of Cohen's kappa. Vach[19] argued that the dependence of Cohen's kappa on the prevalence of the feature of interest "does not matter," because kappa is exactly fulfilling its purpose, which is to improve the interpretation of agreement rates. Indeed, it is intuitive that different populations, regardless of the prevalence of the feature of interest, would yield different kappa values. In fact, the chance correction used in Cohen's kappa actually helps to standardize results across populations, which can be advantageous for comparisons across studies and of the performance of the raters.[20] Rather than criticizing Cohen's kappa for its dependence on the prevalence of features or searching for statistical methods to salvage inefficient studies, the focus should be placed on obtaining populations with a prevalence of the feature of interest near 50%.[21] Nevertheless, one could argue that this is not realistic from a clinical perspective and, furthermore, that doing so hampers the generalizability of the findings to clinical practice.

In contrast, other experts have proposed adjustments and extensions to Cohen's kappa that are suggested to be paradox proof. In addition to Gwet's AC1, alternative measures such as Aickin's α[10] and prevalence- and bias-adjusted kappa[22] have been proposed to address the paradoxical behavior of kappa. One of the most creative alternatives is the B statistic proposed by Bangdiwala and Shankar,[11] which uses a visualization of the agreement between raters and adjusts the observed area of agreement with that expected to result from chance. As revealed in the results, both Aickin's α and B statistic were higher than Cohen's kappa across all five features.

The decision regarding which interrater indices to report should be guided by the purpose of the study, whether reliability or agreement (or both) is of primary interest.[23] Although they are often used interchangeably, there are important differences between the concepts of reliability and agreement.[24,25] In *agreement*, the question of interest is, "Are the ratings identical or close between two pathologists for each case?" In *reliability*, the question of interest is, "How well do the ratings distinguish one case from another?" Hence, agreement indices apply to instruments (or rating criteria) that are used for evaluative purposes, whereas reliability indices are required for instruments that are used for discriminative purposes. Although Cohen's kappa was first proposed to describe agreement between raters, it was argued that, with adjustment of the observed agreement for the chance agreement, an agreement measure can be turned into a reliability measure.[26]

In accordance with the suggestion from Feinstein and Cicchetti,[2,3] we have presented results for positive and negative agreement, in addition to overall agreement. Similar to the concept of sensitivity and specificity in a diagnostic test, these agreement indices distinguish between positive and negative classifications, which may have different implications in clinical practice. A clinical application of positive and negative agreement can be illustrated using the examples of lepidic pattern and acinar pattern from our study. When assessing lepidic pattern, on the basis of the P_{pos} value of 80% and the P_{neg} value of 84%, it may not be necessary to request a second opinion for either an "absent" or a "present" rating. For acinar pattern, however, the P_{pos} and P_{neg} values were 93% and 50%, respectively. This implies that, in practice, if one pathologist rates "absent," it may be worthwhile to obtain the opinion of a second pathologist. In the case of a "present" rating, however, the probability that the second pathologist agrees is 93%, so it is not worthwhile to involve a second pathologist. This reveals the value of specific agreement in clinical practice and that both P_{pos} and P_{neg} are useful contextual metrics in interrater agreement studies.

A recent article by Vach and Gerke[27] conducted a head-to-head comparison of Cohen's kappa and Gwet's AC1. On the basis of the behavior of both metrics under various settings, the study concluded that in the case of no association or maximal disagreement, Gwet's AC1 should not be viewed as a substitute for kappa and that the classification of degrees of agreement in Landis and Koch[6] should not be applied to Gwet's AC1. Even though the extreme scenarios of no association and maximal disagreement are unlikely between pathologists, in the present study, we have argued that agreement studies should present Gwet's AC1 alongside the conventional Cohen's kappa, rather than as a replacement. Much like the convention of presenting both sensitivity and specificity for medical diagnostic tests, the use of multiple indices is based on the acknowledgment that no single index of agreement can be satisfactory for all purposes. In addition, by including complementary indices, such as positive and negative agreement, an interrater study can provide a more clinically relevant determination of interrater variability. With advances in technology, these metrics are readily available in standard statistical software; therefore, researchers are not restricted to reporting only Cohen's kappa.

Agreement statistics depend on feature prevalence. In addition to Cohen's kappa, future interrater variability studies should consider the purpose of the study, report the prevalence of the feature(s) of interest, and include additional agreement statistics such as Gwet's AC1, especially in cases where there is a high prevalence of the feature of interest. In addition to overall agreement, positive and negative agreement should also be reported to allow for clinical and practical interpretation of agreement studies in pathology.

CRediT Authorship Contribution Statement

Kay See Tan: Conceptualization, Data curation, Formal analysis, Methodology, Writing—original draft, Writing—Review, and editing.

Yi-Chen Yeh: Data curation, Formal analysis, Investigation.

Prasad S. Adusumilli: Conceptualization, Investigation, Writing—review and editing.

William D. Travis: Conceptualization, Investigation, Writing—review and editing.

Acknowledgments

This work was supported by the National Institutes of Health/National Cancer Institute Cancer Center Support Grant P30 CA008748 (to Memorial Sloan Kettering Cancer Center). Dr. Adusumilli's laboratory work is supported by grants from the National Institutes of Health (P30 CA008748, R01 CA236615-01, and R01 CA235667), the U.S. Department of Defense (BC132124, LC160212, CA170630, CA180889, and CA200437), the Batishwa Fellowship, the Comedy versus Cancer Award, the Dalle Pezze Foundation, the Derfner Foundation, the Esophageal Cancer Education Fund, the Geoffrey Beene Foundation, the Memorial Sloan Kettering Technology Development Fund, the Miner Fund for Mesothelioma Research, the Mr. William H. Goodwin and Alice Goodwin, the Commonwealth Foundation for Cancer Research, and the Experimental Therapeutics Center of Memorial Sloan Kettering Cancer Center. The sponsors played no role in any aspect of this work.

The authors acknowledge excellent editorial assistance from Christy Rajcoomar and David B. Sewell of Memorial Sloan Kettering Cancer Center.

References

1. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
2. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol*. 1990;43:551-558.
3. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol*. 1990;43:543-549.
4. Yeh YC, Nitadori J, Kadota K, et al. Using frozen section to identify histological patterns in stage I lung adenocarcinoma of ≤3 cm: accuracy and interobserver agreement. *Histopathology*. 2015;66:922-938.
5. 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: executive summary. *Proc Am Thorac Soc*. 2011;8:381-385.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
7. Robitzsch A, Steinfeld J. Item response models for human ratings: overview, estimation methods, and implementation in R. *Psychol Test Assess Model*. 2018;60:101-138.
8. Robitzsch A, Steinfeld J. immer: Item response models for multiple ratings. R package. version 1.1-35; 2018. https://cran.r-project.org/web/packages/immer/index.html. Accessed July 1, 2023.
9. Stevenson M. Evan Sergeant with contributions from Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, epiR: Tools for the Analysis of Epidemiological Data. R package version 2.0.19. https://CRAN.R-project.org/package=epiR. Accessed July 1, 2023.
10. Aickin M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics*. 1990:293-302.
11. Bangdiwala SI, Shankar V. The agreement chart. *BMC Med Res Methodol*. 2013;13:1-7.
12. Meyer D, Zeileis A, Hornik K. vcd: Visualizing Categorical Data. R Package Version 1.4-8. 2020. https://cran.r-project.org/web/packages/vcd/index.html. Accessed July 1, 2023.
13. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008;61:29-48.
14. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:1-7.
15. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213-220.
16. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol Bull*. 1971;76:365.
17. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378.
18. Gwet K. *Handbook of Inter-rater Reliability*. Gaithersburg, MD: STATAXIS Publishing Company; 2001.
19. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. *J Clin Epidemiol*. 2005;58:655-661.
20. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. *J Clin Epidemiol*. 1988;41:959-968.
21. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *J Clin Epidemiol*. 2000;53:499-503.

22. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46:423–429.
23. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int J Nurs Stud*. 2011;48:661–671.
24. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59:1033–1039.
25. Kottner J, Streiner DL. The difference between reliability and agreement. *J Clin Epidemiol*. 2011;64: 701–702.
26. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*. 1987;40:171–178.
27. Vach W, Gerke O. Gwet’s AC1 is not a substitute for Cohen’s kappa—a comparison of basic properties. *MethodsX*. 2023;10:102212.