Estimating Interobserver Reliability

To the Editor

When 3 pathologists examine a set of slides for the presence or absence of a characteristic, a reasonable estimate of interobserver reliability can be derived by simple mathematical manipulation of a known equation:1,2

$$\kappa = 1 - \frac{dis \times 3n}{[pos \times (3n - pos)]},$$

where \(n\) is the number of cases, \(dis\) is the number of cases which lack uniform agreement, and \(pos\) is the number of positive ratings. The application is suitable for studies of a moderate size and produces results similar to those of complex formulæ that require computer programs.

A recent article3 demonstrates both the utility and potential utility of the shortcut application. In that article, 29 smears of adenocarcinoma-in-situ (AIS) and benign AIS mimics were examined by 3 raters, so \(n = 29\). Six AIS cases and 4 mimics did not receive uniform ratings, so \(dis = 10\). Of 51 ratings of 17 AIS cases, 8 were wrong; of 36 ratings of 12 benign cases, 4 were errors. So \(pos = 51 - 8 + 4 = 47\). Thus,

$$\kappa = 1 - \frac{10 \times 3 \times 29}{[47 \times (3 \times 29 - 47)]} = .54.$$  

But what does a \(\kappa\) of .54 mean? Landis and Koch declared \(\kappa\) values of .41–.6 moderate.4 Fleiss said a hypothetical set of 25 subjects having a \(\kappa\) of .54 had “only a modest degree of interrater agreement.”2 More to the point, Farmer et al.5 in a study of melanocytic lesions, said a \(\kappa\) of .50 showed better criteria are needed. Thus, AIS is a diagnosis experts cannot agree upon; missing a case represents allowable error in judgment.

One might say these were not classic AIS cases, but such a canard runs afool of several arguments. First, the cases come from a teaching file at least 1 author has had access to for many years, one with which he has greater familiarity than most do with slides they have not seen. Second, the authors are more adept, and know one another’s diagnostic styles better, than most pathologists. Third, many people tell me this is the most common pattern of the rare AIS cases they see: AIS, in their experience, does not usually have large cells. Fourth, the authors knew the diagnostic possibilities, discussed diagnostic criteria, and examined sample cases before examining the slides in question, things possible in a study, but impossible in practice. Finally, even if AIS is rare, the benign mimics delineated in this article are common; it is vital to be able to say if the best of us can reliably separate these common benign cell groups from malignant processes. Indeed, the last reason is the most important because a study comparing AIS to, say, simple reactive endocervical cells would tell us nothing. Perhaps a \(\kappa\) of .54 is a good estimate concerning AIS in general of the best to be expected of pathologists who do not know one another, have no advance knowledge of the differential diagnoses, and have not discussed criteria and sample cases before examining their slides.

The authors list and statistically test 15 consensus criteria. Since no adjustment for multiple comparisons is given, the interested reader might provide one of his or her own choosing. Using Bonferroni’s rule,6 \(a = .05/15 = .0033\), 2 tests show results unexplainable by chance. Thus, when 3 experts examine a slide with this differential diagnosis and at least 2 see coarse chromatin or less than 3 fragments of

References

endometrial stroma, there is a statistically significant association with a histologic diagnosis of AIS. On a negative basis, $\kappa$ values are needed because they help protect patients from overly exuberant statistics; hard, personal experience shows that criteria with, initially, very low $P$ values may have even more abysmal $\kappa$ values than .54.2 On a positive basis, one longs for $\kappa$ values of these criteria because the authors are excellent pathologists and the criteria are not only used for AIS, but for many other diagnoses. Since $n$ is constant at 29, one need simply count the number of positive ratings and the number of cases with non-uniform ratings to calculate $\kappa$. The $\kappa$ values are valid for all criteria except those seen in only a few specimens, such as prominent nucleoli and irregular nuclei.

In this regard, the American Society of Clinical Pathologists may have a special role. On its Web site, the ASCP could post daily photographs with an arrow to a particular cell or cell group. The viewers, who could be required to be board certified, would be questioned whether they thought the object had a certain characteristic, such as degenerated chromatin. Since $\kappa$ tests require constancy neither of raters nor even of number of raters, a large number of raters of a large number of cells could be collected for extremely precise $\kappa$ values for each criterion.

Although reproducible diagnoses do not always require reproducible criteria, reliable criteria surely help ensure accurate, reproducible diagnoses. Moreover, learning which criteria are unreliable may lead us to the source of irreproducible diagnoses. Criteria are the building materials of diagnoses; it is vital to learn whether our diagnostic schemas are made of brick or of straw. What are the $\kappa$ values for this study’s criteria?

The Authors’ Reply

We thank Dr Wachtel for his interest and observations concerning our study. We did not publish $\kappa$ values for interobserver variability of the diagnoses, although we discussed doing so, because the cases themselves were selected for their difficulty and do not reflect the more common presentations of either AIS or its mimics. Since a prior study indicated that the small cell “endometrioid” cell pattern of AIS had been a source of falsely negative diagnoses,1 this follow-up study2 was conceived to see if we could establish criteria to separate this cytologic presentation of AIS from its more commonly encountered mimics, rather than to compare individual diagnoses based on these criteria. We are currently conducting a follow-up study using multiple observers of varying experience levels to test the diagnostic accuracy of this differential diagnosis using the criteria we have established.

Although we indicate how our criteria were established and correlated with the correct diagnosis in our methods section, we did not consider publishing the interobserver $\kappa$ values for each criterion. Although Dr Wachtel is correct when he says that reliable criteria help ensure accurate diagnoses, cytologists are aware that a diagnosis depends upon judgment and experience in simultaneously interpreting and weighing many criteria. Thus, even if the $\kappa$ values for 1 or another criterion that statistically correlated with a diagnosis in our study were rated as only fair or even poor, that does not negate the potential usefulness of the criterion. Indeed, although none of the 3 study participants had seen any of the unknown slides beforehand, one of us was able to diagnose 27 of the 29 cases correctly, indicating that it is possible in most cases to do so.

References