Kappa statistics as indicators of quality assurance in histopathology and cytopathology

S S Cross

Abstract

Kappa statistics are widely-used to assess performance in quality assurance schemes. Low values, however, are difficult to interpret, especially when confidence intervals have not been calculated. A model of a dichotomous decision in pathology (benignancy or malignancy in fine needle aspirates of the breast) was used to calculate kappa statistics (with confidence limits) for increasing false positive rates. It was found that the level at which the upper 95% confidence interval for the kappa statistic fell below 1 was an insensitive method of detecting unsatisfactory performance as at that level the false positive rate was unacceptable high (>1%) for all populations of specimens less than 800 in number. Either large populations of samples are required in quality assurance schemes which use kappa statistics (which may well be impractical) or

other methods of assessing performance, possibly with weighted outcomes, are required. (J Clin Pathol 1996;49:597–599)

Keywords: quality assurance schemes, kappa statistics, cytopathology.

Internal and external quality assurance schemes, proficiency testing and related exercises are seen as useful ways of promoting uniform high quality reporting of specimens in histopathology and cytopathology. Assessment of results in these schemes differ from other pathological specialties, such as clinical chemistry and haematology, in that the results are discrete diagnostic categories (for example, severe dyskaryosis in a cervical smear) rather than variable parameters (for example, haemoglobin) and for this reason kappa statistics are often used as indicators of performance.\textsuperscript{1-3}

Kappa statistics measure levels of agreement between two observers and make allowance for the degree of agreement that would occur by chance alone. In histopathology and cytopathology there has been discussion about the most appropriate reference diagnosis to be used in the calculation of kappa statistics (for example, the consensus of participants in the scheme, an expert diagnosis, a diagnosis verified by non-histopathological criteria, etc.) but another important factor to be considered is the significance of any calculated kappa statistic in its context. In almost all quality assurance schemes which use kappa statistics the statistic is given as a single figure for each diagnosis, without confidence limits, and it is difficult to assess the significance of low values.\textsuperscript{4} In the present study a mathematical model of a simple quality assurance scheme is used with calculation of kappa statistics with confidence limits and the significance of kappa statistic values is discussed in relation to other test performance statistics.

Methods
A statistical model was developed which simulated reporting of breast fine needle aspirates with a dichotomous outcome (benign/malignant). Populations from 50 to 1600 specimens were used and for increasing numbers of false positive results, the kappa statistics were calculated with 95% confidence intervals (CI) using the method described by Silcocks.\textsuperscript{5} Details of the model and calculations are given in the appendix. Other performance parameters, including the positive predictive value (PPV) of the test, were also calculated. The number of false positive results sufficient for the calculated upper 95% CI to fall below 1 was found for each population size. Two systems with differing mixes of benign and malignant specimens were used; a 50% benign, 50% malignant mix which could be used in a quality assurance scheme and a 80% benign, 20% malignant mix which represents the prior probabilities in breast fine needle aspirates received in the Department of Pathology at the Royal Hallamshire Hospital.

Table 1 Results for populations of 50% benign, 50% malignant specimens

<table>
<thead>
<tr>
<th>No. of cases</th>
<th>False positive results*</th>
<th>Kappa</th>
<th>Lower CI</th>
<th>Upper CI</th>
<th>PPV†</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4</td>
<td>0.78</td>
<td>0.58</td>
<td>0.99</td>
<td>71.4</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.86</td>
<td>0.74</td>
<td>0.98</td>
<td>80.0</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>0.90</td>
<td>0.81</td>
<td>0.99</td>
<td>88.9</td>
</tr>
<tr>
<td>400</td>
<td>6</td>
<td>0.95</td>
<td>0.92</td>
<td>0.99</td>
<td>93.0</td>
</tr>
<tr>
<td>800</td>
<td>6</td>
<td>0.97</td>
<td>0.95</td>
<td>0.99</td>
<td>96.4</td>
</tr>
<tr>
<td>1600</td>
<td>9</td>
<td>0.98</td>
<td>0.97</td>
<td>0.99</td>
<td>97.3</td>
</tr>
</tbody>
</table>

*Number of false positive results required for the upper 95% CI of the kappa statistic to fall below 1.
†PPV of a malignant result.

Table 2 Results for populations of 80% benign, 20% malignant specimens

<table>
<thead>
<tr>
<th>No. of cases</th>
<th>False positive results*</th>
<th>Kappa</th>
<th>Lower CI</th>
<th>Upper CI</th>
<th>PPV†</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4</td>
<td>0.84</td>
<td>0.63</td>
<td>0.97</td>
<td>83.3</td>
</tr>
<tr>
<td>100</td>
<td>5</td>
<td>0.90</td>
<td>0.81</td>
<td>0.99</td>
<td>90.9</td>
</tr>
<tr>
<td>200</td>
<td>5</td>
<td>0.95</td>
<td>0.91</td>
<td>0.99</td>
<td>95.2</td>
</tr>
<tr>
<td>400</td>
<td>6</td>
<td>0.97</td>
<td>0.95</td>
<td>0.99</td>
<td>97.0</td>
</tr>
<tr>
<td>800</td>
<td>8</td>
<td>0.98</td>
<td>0.97</td>
<td>0.99</td>
<td>98.0</td>
</tr>
<tr>
<td>1600</td>
<td>11</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>98.6</td>
</tr>
</tbody>
</table>

*Number of false positive results required for the upper 95% CI of the kappa statistic to fall below 1.
†PPV of a malignant result.

Results and Discussion
Tables 1 and 2 summarise the results.
If the kappa value is less than one, then we are 100% sure that performance is less than perfect and the upper confidence limit must also be less than one. In practice, however, methods of calculating confidence limits are approximate and high kappa values do have calculated upper limits of one when standard methods are used. Thus, to show by calculation that performance is significantly (p < 0.05) below a very high level (close to perfection) the calculated upper confidence limit must be less than one. At all population sizes and mixtures in the present study, however, the number of false positive results required to produce an upper confidence limit below one was unacceptably high. The Cytology Subgroup of the National Coordinating Committee for Breast Screening Pathology recommend that the false positive rate should be less than 1%, but in this study the upper confidence limit of the kappa statistic fell below one, with a false positive rate of less than 1% only in populations with 800 and 1600 cases in a 50:50 population mix (rows 5 and 6 of table 1) and in the population of 1600 cases in the 80:20 mix (row 6 of table 2). Population numbers of 800 and 1600 represent the number of fine needle aspirates of the breast a pathologist might see during one or two years in a busy laboratory rather than the number of specimens that could be included in a quality assurance scheme.

The model used in this study is the simplest that could occur in pathology with a dichotomous outcome that can be verified (by clinical examination, mammography, histology, and follow up). Most other diagnostic situations in pathology would involve more diagnostic categories and the outcomes would be less easily verified (for example, diagnosis of inflammatory bowel disease by colorectal biopsy\textsuperscript{7} or assessment of chronic gastritis\textsuperscript{8}). A
greater number of diagnostic outcomes would produce wider confidence limits for the kappa statistics if the population size remained the same. It therefore seems that the kappa statistic is not a sensitive indicator of performance in cytopathological or histopathological quality assurance schemes unless large numbers of samples are used. Other indicators should be considered, possibly using systems which weight the outcomes according to the importance of the result in clinical practice. The present study did not include the function of time and the persistence of low kappa statistics over a number of cycles of a quality assurance scheme would be a more reliable indicator of possible unsatisfactory performance. However, the frequency of cycles in most schemes is such that at least two years would elapse before such an assessment would be possible.

1 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

In the initial state of the model \( b = 0 \) and \( c = 0 \), and the rest of the cases are split between \( a \) and \( d \) in the proportions for that particular model (either 50:50 or 80:20 as described in the text). For each iteration of the model \( b = b + 1 \) and \( a = a - 1 \)—that is, there is replacement of one true negative with a false positive result. For each iteration of the model the parameters given below are calculated.

\[
\text{Observed probability } P_b = \frac{a + d}{i}
\]

\[
\text{Expected probability } P_e = \frac{\frac{g + h}{i} - i}{i}
\]

\[
\kappa = \frac{P_b - P_e}{1 - P_e}
\]

\[
\text{Standard error of } \kappa = \sqrt{\frac{P_b(1-P_b)}{i(1-P_e)^2}}
\]

95% CI

\[
CI = \kappa \pm (1.96* \text{s.e.} \kappa)
\]

PPV of a malignant diagnosis

\[
\text{PPV} = \frac{d}{d + b}
\]

Appendix

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Diagnosis</th>
<th>Benign</th>
<th>Malignant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>a</td>
<td>b</td>
<td>g</td>
<td></td>
</tr>
<tr>
<td>Malignant</td>
<td>c</td>
<td>d</td>
<td>h</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>e</td>
<td>f</td>
<td>i</td>
<td></td>
</tr>
</tbody>
</table>

Malignant lymphoma in congenital dyserythropoietic anaemia type III

W G McCluggage, D Hull, E Mayne, H Bharucha, S N Wickramasinghe

Abstract

A 60 year old woman with congenital dyserythropoietic anaemia (CDA) type III developed a malignant T cell lymphoma with cutaneous and widespread nodal involvement. Bone marrow aspirates showed erythroid hyperplasia and dyserythropoiesis with multinucleate erythroblasts and giant cells, in keeping with CDA type III. Electron microscopy showed multinucleate erythroblasts with notably irregular nuclear outlines and intranuclear clefts. The development of malignant lymphoma in this patient, together with a documented high prevalence of monoclonal gammapathy and multiple myeloma and a single case of Hodgkin's disease, may indicate an increased incidence of lymphoproliferative disease in CDA type III.

Keywords: congenital dyserythropoietic anaemia type III, malignant lymphoma, electron microscopy.