Reproducibility of Malignant Pleural Mesothelioma Histopathologic Subtyping: Methodologic Issue

To the Editor.—We were interested to read the paper by Brcic and colleagues in the Archives of Pathology & Laboratory Medicine. The authors aimed to determine the interobserver and intraobserver reproducibility in the histologic differentiation among the main types of malignant pleural mesothelioma (MPM), and in further subtyping of epithelioid MPM. One representative hematoxylin–eosin–stained slide was selected from the archive for each of 200 patients with MPM. The slides were reviewed independently by 3 pathologists and classified according to the current World Health Organization classification of pleural tumors. The interobserver and intraobserver agreement was interpreted using the Cohen k statistic. Based on their results, the overall interobserver agreement for histologic subtyping of mesothelioma was fair (k = 0.36) and the agreement was increased to substantial (k = 0.63) in the second round; therefore, improvement was found in interobserver agreement for all types of MPM and for most epithelioid subtypes.

Knowing that there is no value of k that is internationally a sign of good agreement is of great importance. The k value to assess the agreement of a qualitative variable has 2 weaknesses, as follows. First, it depends on the prevalence in each category. In other words, it is possible to have different k values with the same percentage for both concordant and discordant cells! In the Table, in both situations (a) and (b), concordant (agreement) and discordant (disagreement) cells have prevalences of 90% and 10%, respectively; however, we get different k values (0.44, moderate, and 0.80, very good). The k value is also dependent on the number of categories. In such a situation, especially having more than 2 observers, our suggestion is to apply weighted or Fleiss k because the mentioned estimates provide us unbiased results.

Brcic et al concluded that the moderate to substantial agreement in histologic typing and subtyping of MPM can be achieved. Such a conclusion should be supported by the above-mentioned statistical and methodologic issues.

In this letter, the limitations of the k value to assess reliability are mentioned.

Mehdi Naderi, MSc; Siamak Sabour, MD, MSc, DSc, PhD

1 School of Paramedical, Kermanshah University of Medical Sciences, Kermanshah, Iran; 2 Department of Clinical Epidemiology, School of Health, Safety Promotion and Injury Prevention Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.


Accepted for publication May 7, 2018.

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


In Reply.—We are grateful for the opportunity to answer issues raised in a letter concerning methodology in our article analyzing interobserver and intraobserver reproducibility in histologic subtyping of malignant pleural mesothelioma. First of all, we have to correct an error in the Materials and Methods section of our article: instead of the Cohen k (as it was stated) we have actually used the Fleiss k as a measure of agreement. This error occurred during our own editing process of the article, and was later overlooked. We deeply regret the error and appreciate this letter. Furthermore, we would also like to bring up some points about k values.

Since its publication in 1960, as a chance corrected measure of agreement, Cohen k has prompted a vast amount of articles about its properties. The dependence of k on prevalence arises from chance correction, which has to be greater if one category is more prevalent. Alternatively, one can state that it is more difficult to achieve high agreement in a very homogeneous population, and therefore k is lower, although the proportion of discordant ratings is the same.

In the letter, two combined 2 × 2 tables of ratings with the same fraction of discordant ratings, but different prevalences of positive ratings, are given for which Cohen k is different. However, for symmetrical tables, the Fleiss k is the same as Cohen k, thus it is affected by the same alleged weakness.

Limitation of k for Comparison of 2 Pathologists’ Diagnoses With Different Prevalence in the 2 Categories

<table>
<thead>
<tr>
<th>Pathologist 2</th>
<th>Positive Result</th>
<th>Negative Result</th>
<th>Total, %</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation (a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive result</td>
<td>85</td>
<td>5</td>
<td>90</td>
<td>0.44 (moderate)</td>
</tr>
<tr>
<td>Negative result</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total, %</td>
<td>90</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Situation (b)</td>
<td></td>
<td></td>
<td></td>
<td>0.80 (very good)</td>
</tr>
<tr>
<td>Positive result</td>
<td>45</td>
<td>5</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Negative result</td>
<td>5</td>
<td>45</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Total, %</td>
<td>50</td>
<td></td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

* Authors’ own hypothetical data to show limitation of k value to assess reliability.
If all 10 discordances in the above-mentioned combined table would be of the type that pathologist 1 rates positively and pathologist 2 rates negatively, and all concordant ratings are left unchanged, Cohen $\kappa$ would be increased, whereas Fleiss $\kappa$ would remain unchanged. The effect can be quite substantial, raising doubts about the adequacy of Cohen $\kappa$ as a measure of rater agreement. Fleiss $\kappa$ does not have that flaw; however, its shortcoming is that Fleiss $\kappa$ is not necessarily zero under rater independence. (For example, the table with 2 pathologists rating negative/negative and positive/positive 4 times each, negative/positive 2 times, and positive/negative 8 times has independent ratings, thus Cohen $\kappa$ is zero. Fleiss $\kappa$ is 0.111, however.)

We conclude that $\kappa$ type measures of agreement depend on prevalence by their very nature, and that Fleiss $\kappa$ is unaffected by an asymmetric distribution of discordances.

Luka Brcic, MD, PhD; Gregor Vlacic, MD; Franz Quehenberger, PhD; Izidor Kern, MD

1 Institute of Pathology and 3 Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria; 2 Cytology and Pathology Laboratory, University Clinic of Respiratory and Allergic Diseases, Golnik, Slovenia.


Letters to the Editor

Use of a Web-Based Checklist to Improve Compliance With Medicare Access and CHIP Reauthorization Act of 2015 Reporting

To the Editor.—Checklists can improve compliance with defined standards in a variety of situations,1,2 including surgical pathology.3–6 Along these lines, we have previously shown that a Web-based checklist can improve the completeness of synoptic reporting in surgical pathology.7

The Medicare Access and CHIP [Children's Health Insurance Program] Reauthorization Act of 2015 (MACRA) for pathology is a standard that requires that pathologists report both the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10) and billing codes for specific types of pathology specimens. In the past and without the use of a checklist, the accurate reporting of the MACRA features in our laboratory was relatively low (1297 of 1707; 76%). We wondered if the inclusion of a question on our Web-based synoptic report checklist related to MACRA compliance would improve our laboratory’s performance.

Since August 2016, our synoptic reports for breast, lung, prostate, and melanoma resections have included a question that asks the pathologist to report both the ICD10 and the billing code for that specimen. In contrast, there was no checklist for reporting these features on esophageal biopsies for Barrett mucosa. For the 19 months ending February 28, 2018, compliance for reports using the Web-based checklist has been significantly better (100%; 1667 of 1667) compared with that for Barrett mucosa (82.9%; 169 of 204; $P < .001$, 2-tailed $\chi^2$ test). These results demonstrate that checklists can be of value in improving compliance in the reporting of surgical pathology for topics other than synoptic reporting. The exact benefit of any particular checklist varies with the specifics of how it is implemented.8 We have found that reminder functions are critical to achieving a very high rate of compliance. Force functions such as those seen in the College of American Pathologists (CAP) electronic cancer checklist would likely also achieve a very high rate of compliance. We would encourage pathologists and the CAP to consider incorporating additional features into their own synoptic report checklists for surgical pathology.

Andrew A. Renshaw, MD; Edwin W. Gould, MD

Department of Pathology, Baptist Hospital and Miami Cancer Institute, Miami, Florida


Received for publication June 14, 2018.

1312 Arch Pathol Lab Med—Vol 142, November 2018