Agreement and Error Rates Using Blinded Review to Evaluate Surgical Pathology of Biopsy Material

Andrew A. Renshaw, MD,1 Norberto Cartagena, MD,1 Scott R. Granter, MD,2 and Edwin W. Gould, MD1

Abstract

Blinded review has been shown to be an excellent method to detect disagreements and errors and improve performance in gynecologic cytology. Preliminary studies suggest it may be valuable in surgical pathology.

We reviewed 5,000 sequential outpatient surgical pathology biopsy cases without knowledge of the original diagnosis or history and compared the results with those of the original diagnosis.

Complete agreement was obtained in 91.12% of cases. The technique of blinded review of surgical pathology biopsy material had a sensitivity of more than 99%, failing to identify an abnormality in 19 cases. Although there was a significant level of diagnostic disagreement (444 cases), primarily due to differences in diagnostic thresholds (292 cases), diagnoses that resulted in a change in the original report (true errors) were present in only 5 cases, and only 4 were clinically significant. This clinically significant error rate of 0.08% is significantly lower than previously published error rates.

Blinded review is a sensitive (99%) and effective method to identify areas of disagreement and errors in surgical pathology biopsy material. The relatively high rate of disagreement found with blinded review coupled with the very low rate of error highlights the substantial potential for bias in nonblinded reviews.

Diagnostic accuracy in surgical pathology is critical for appropriate patient care, measuring quality, and medicolegal evaluation. Although diagnostic agreement has been assessed in cytopathology and surgical pathology,1-11 many of these studies did not distinguish between diagnostic agreement and accuracy and were not blinded with respect to the initial diagnosis. The potential bias of nonblinded review has been studied extensively in gynecologic cytology, and blinded review, which may be less biased than nonblinded review, is known to be an effective method to measure performance in cytopathology.12-17 Recent preliminary reports suggest that the same technique is sensitive and effective for evaluating surgical pathology material.18 We sought to evaluate this technique in a large series of outpatient surgical pathology biopsy material.

Materials and Methods

All outpatient surgical pathology biopsy specimens received at the Baptist Hospital of Miami, Miami, FL, from September 2001 to June 2002 (with the exception of approximately 3 weeks of vacation) and not interpreted by the reviewer were reviewed by one of us (A.A.R.) without knowledge of the initial diagnosis or clinical history. Type and number of disagreements were determined as outlined previously.19 All cases were reviewed within 48 hours of the original diagnosis and usually within 12 hours of the original diagnosis.

Disagreements were defined as follows19: (1) A threshold disagreement is concurrence on the nature of the lesion but disagreement about its degree. Common examples
include disagreement between actinic keratosis and squamous cell carcinoma in situ or between mild and moderate atypia in a dysplastic nevus. (2) Disagreements related to process type were designated *type* disagreements. For example, a type disagreement would include a difference of opinion about whether a tumor is best classified as basal cell carcinoma or squamous cell carcinoma. In some cases, the categorization of a disagreement into either threshold or type may be subjective. All categorizations in the study ultimately were made by one of us (A.A.R.). The *false-negative rate* is related to screening and measures whether a lesion is recognized at all. *False-positive diagnoses* were diagnoses that were made in cases in which the lesion was not present.

For a case to be classified as an error rather than simply a disagreement, all disagreements that were thought to potentially be errors were reviewed by the initial pathologist and the reviewer together. If there was agreement that an error was present, this qualified as a diagnostic error only if the error was important enough to result in a change in the initial diagnostic report. If agreement could not be reached but either pathologist thought an error still might be present, the case was reviewed by an outside pathologist (S.R.G.) whose diagnosis was accepted as the “gold standard.” Potential clinically significant errors were those in which treatment or prognosis might be different.

Statistical comparisons of categorical data were performed using a two-tailed $\chi^2$ test.

**Results**

The distribution of biopsy material by site of origin is detailed in Table 1. An abnormality was identified in a total of 4,922 cases (98.44%).

The number and type of disagreements identified are listed in Table 2. Discrepancies were identified in a total of 444 cases (8.88%).

Blinded screening had an excellent sensitivity, missing the lesion in only 19 cases. Since there was a diagnostic abnormality in 4,922 cases, the sensitivity of blinded rescreening was 99.61% (4,903/4,922).

**Table 1**

<table>
<thead>
<tr>
<th>Site</th>
<th>No. of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skin</td>
<td>4,631</td>
</tr>
<tr>
<td>Gynecologic tract</td>
<td>223</td>
</tr>
<tr>
<td>Bone marrow</td>
<td>93</td>
</tr>
<tr>
<td>Gastrointestinal tract</td>
<td>32</td>
</tr>
<tr>
<td>Breast</td>
<td>17</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>5,000</td>
</tr>
</tbody>
</table>

The sensitivity of the original screening was extremely high. Although there were 8 lesions identified on review that were missed in the initial diagnosis, all of which were agreed on by both reviewers, in only 1 case was the misdiagnosis deemed sufficiently important to alter the original pathologic diagnosis.

Differences in diagnostic thresholds were very common, accounting for 292 disagreements.

Although there were a total of 444 cases in which a disagreement was identified, this disagreement resulted in a change in the original pathology report in only 5 cases (0.10%). These included 1 false-negative error, 2 errors of type, 1 error of threshold, and 1 clerical error (Table 3). All except the clerical error were clinically significant. This clinically significant error rate of 4 of 5,000 cases, or 0.08%, is significantly lower than the lowest previously published clinically significant error rate of 0.3% ($P = .02$).

**Discussion**

The present study differs from other studies on diagnostic agreement in 3 ways. First, it uses blinded review. Although blinded review has been used in gynecologic cytology and in preliminary studies in surgical pathology, this report represents the largest surgical pathology series that has used the method. Although there are limitations to this technique, these are not as great as one might imagine. While there certainly were cases that were misclassified owing to a lack of clinical history, the number of such cases was relatively low. In addition, the sensitivity of the technique (>99%) was quite high. Both of these results suggest that the method is effective for identifying areas of disagreement. Indeed, the very high rate of threshold disagreements (almost 6%) strongly suggests that there are many biopsy specimens that can be diagnosed differently by different
The nature of the errors in the present study is of interest. While some cases, such as the case of herpes, represented difficult screening challenges, others, such as the misidentification of proliferative endometrium as secretory endometrium, did not. When these cases were brought to the original pathologist’s attention, there was no disagreement about the correct diagnosis. Seen out of context, such errors seem sophomoric. Indeed, one opinion (A.A.R.) after performing this study is that while screening error remains a significant source of potential diagnostic error, a potentially more significant source of error is when a pathologist simply goes in the wrong direction without any clear reason. Practicing surgical pathologists routinely and rapidly process literally thousands of images a day, and it may be that for whatever reason—clinical history, distraction, or human nature—every once in a while an error is made for which there is no good explanation.

Whether blinded review should be incorporated into the routine evaluation of surgical pathology material as a method of quality improvement or as a method for documenting quality is not clear. The present study involved review of 5,000 cases, took 10 months, and resulted in the correction of only 4 clinically significant errors. While it is satisfying to demonstrate that the error rate in our laboratory is this low and also to believe that with blinded review the error rate approaches zero, the time and effort involved may limit routine implementation of blinded review. For example, if one wished to show a significantly lower error rate than we found, one would have to identify no errors in a review of almost 7,000 cases. If more complex cases were reviewed, the time involved in the review obviously would increase.

We have shown that blinded review is a sensitive (>99%) and effective method to identify areas of disagreement and errors in surgical pathology biopsy material. The relatively high rate of disagreement found with blinded review coupled with the very low rate of error highlights the significant potential for bias in nonblinded reviews. In other words, this low error rate might easily be increased by examining cases that only represent disagreements in a biased light.

---

### Table 3

**Diagnostic Errors Identified by Blinded Review**

<table>
<thead>
<tr>
<th>Original Diagnosis</th>
<th>Final Diagnosis</th>
<th>Type of Error</th>
<th>Clinically Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ulcer and granulation tissue</td>
<td>Herpes</td>
<td>False-negative</td>
<td>Yes</td>
</tr>
<tr>
<td>Benign verrucous acanthoma</td>
<td>Favor squamous cell carcinoma, keratoacanthoma type</td>
<td>Threshold</td>
<td>Yes</td>
</tr>
<tr>
<td>Inactive endometrium</td>
<td>Weakly secretory endometrium</td>
<td>Type</td>
<td>Yes</td>
</tr>
<tr>
<td>Secretory endometrium, day 22</td>
<td>Proliferative endometrium</td>
<td>Type</td>
<td>Yes</td>
</tr>
<tr>
<td>Dermal nevus</td>
<td>Seborrheic keratosis</td>
<td>Clerical error</td>
<td>No</td>
</tr>
<tr>
<td>Benign verrucous acanthoma</td>
<td>Favor squamous cell carcinoma, keratoacanthoma type</td>
<td>Threshold</td>
<td>Yes</td>
</tr>
<tr>
<td>Inactive endometrium</td>
<td>Weakly secretory endometrium</td>
<td>Type</td>
<td>Yes</td>
</tr>
<tr>
<td>Secretory endometrium, day 22</td>
<td>Proliferative endometrium</td>
<td>Type</td>
<td>Yes</td>
</tr>
<tr>
<td>Dermal nevus</td>
<td>Seborrheic keratosis</td>
<td>Clerical error</td>
<td>No</td>
</tr>
</tbody>
</table>

---

observers. Fortunately, most of these disagreements were not thought to be important enough to warrant a change in the original pathology report. However, a disagreement rate much higher than in any previous studies, none of which used blinded review, suggests that nonblinded review may underestimate the true rate of potential disagreement.

The second way in which this study differs from previous studies is in how we defined error as distinct from disagreement. In most cases, there is no separate gold standard with which to evaluate a diagnosis. To resolve this issue, we have taken a very pragmatic approach to defining errors. Only cases in which there was a change to the original report qualified as errors. Obviously, the timely performance of the review makes the alteration of the initial report more practical and less potentially litigious. This definition of error may result in an apparently reduced error rate. For example, although there were 8 cases in which an additional lesion was identified by the review, and in every case the initial pathologist agreed the lesion was present, in only 1 case was a change in the initial report made. Nevertheless, in any case with potential clinical significance, a third reviewer always was available, and in every case in which a change in the pathology report would have had clinical significance, the change was made. This suggests that the clinically significant error rate we have demonstrated is comparable to those in previous studies.

Third, the nature of the material used in this study is somewhat different from that of previous studies. The present study primarily used outpatient material and was heavily weighted toward dermatopathology. The nature, degree, and type of errors found in this material may be different from that found in studies that used a higher percentage of material from other anatomic sites.

Nevertheless, the results of this study are reassuring. Previous well-publicized studies have suggested that the error rate in surgical pathology may be much higher than we have shown. However, these studies come from consultation material, and the bias in this material may have affected these results. Our study suggests that the “true” error rate in unselected surgical biopsy material is much lower, approximately 1 in 1,000 cases in this series.
From the Departments of Pathology, 1Baptist Hospital of Miami, Miami, FL, and 2Brigham & Women’s Hospital and Harvard Medical School, Boston, MA.

Address reprint requests to Dr Renshaw: Dept of Pathology, Baptist Hospital of Miami, 8900 N Kendall Dr, Miami, FL 33176.

References


