

Blinded Review of Papanicolaou Smears in the Context of Litigation

Using Statistical Analysis to Define Appropriate Thresholds

Andrew A. Renshaw, M.D.¹
Mary L. Young, M.S.²
E. Blair Holladay, Ph.D.³

¹ Department of Pathology, Baptist Hospital of Miami, Miami, Florida.

² Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina.

³ Center for Quality Improvement in GYN Cytopathology, Medical University of South Carolina, Charleston, South Carolina.

BACKGROUND. Blinded review has been endorsed by several cytology and pathology organizations as the most appropriate method for the review of cervicovaginal specimens in the context of litigation. Methods for determining the statistical validity of this method were evaluated.

METHODS. First, the authors calculated the sample size needed to detect various differences in case difficulty or ease of interpretation, in which ease of interpretation is defined as the percentage of the time a case could be identified as abnormal by routine screening. Very easy cases could be identified most reliably, whereas more difficult cases were detected less regularly and less reliably. Using construct sample sizes, the authors calculated the number of abnormal reviews that may be helpful to conclude that the case's difficulty or ease of interpretation was statistically significantly different from another case of a specified difficulty. Finally, they examined the effect of using two separate cutoff parameters to make these distinctions.

RESULTS. Depending on the threshold chosen, it was determined that improvements in the statistically meaningful distinctions may be made using 15 or 20 reviews. To distinguish between routine false-negative cases (ease of detection, 40%) and routine cases (ease of detection, 80%), the thresholds may be set at 5 of 10 reviews (a case that would not be detected regularly and reliably in any given laboratory) and 7 of 10 reviews (defining a case that would be identified regularly and reliably), respectively.

CONCLUSIONS. The authors provide data that can be used to interpret the results of a blinded review in a statistically appropriate manner. To improve the utility of blinded reviews, the standards are defined explicitly. *Cancer (Cancer Cytopathol)* 2004;102:136–41. © 2004 American Cancer Society.

KEYWORDS: pathology, cytopathology, accuracy, litigation, blinded review, sensitivity, quality.

The review of cervicovaginal specimens in cytology laboratories is not perfect and is associated with multiple known error rates.¹ In studies that defined error as any abnormality (atypical squamous cells [ASC] or ASC of undetermined significance and worse) and used a second review of the specimen as the gold standard, the average sensitivity (the ability to reliably recognize a specimen as abnormal) of the routine review of these specimens in large clinical studies was reported to be approximately 70–80%.¹ This sensitivity is not limited to ASC; similar results were obtained when the threshold of low-grade squamous intraepithelial lesion was used.¹ Despite this known, relatively large error rate, Papanicolaou (Pap) smears with missed abnormalities remain a significant source of litigation in our society. Re-

See related editorial on pages 133–5, this issue.

Address for reprints: Andrew A. Renshaw, M.D., Department of Pathology, Baptist Hospital of Miami, 8900 N. Kendall Drive, Miami, FL 33176; Fax: (305) 598-5986; E-mail: andrewr@bhssf.org

Received July 15, 2003; accepted January 12, 2004.

TABLE 1
Probability of Specific Results (the Number of Abnormal Reviews) in Blinded Review for Cases with Known Ease of Interpretation

Result	Ease of interpretation of case							
	20%	40%	60%	80%	90%	95%	99%	99.9%
0/10	0.11	0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
1/10	0.27	0.04	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
2/10	0.30	0.12	0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
3/10	0.20	0.21	0.04	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
4/10	0.09	0.25	0.11	0.01	< 0.01	< 0.01	< 0.01	< 0.01
5/10	0.03	0.20	0.20	0.03	< 0.01	< 0.01	< 0.01	< 0.01
6/10	0.01	0.11	0.25	0.09	0.01	< 0.01	< 0.01	< 0.01
7/10	< 0.01	0.04	0.21	0.20	0.06	0.01	< 0.01	< 0.01
8/10	< 0.01	0.01	0.12	0.30	0.19	0.07	< 0.01	< 0.01
9/10	< 0.01	< 0.01	0.04	0.27	0.39	0.31	0.09	< 0.01
10/10	< 0.01	< 0.01	0.01	0.11	0.35	0.60	0.90	0.99

cently, blinded review of Pap smears in the context of litigation has become the standard litmus test, and multiple national and regional cytopathology societies have endorsed this as the most appropriate method with which to distinguish acceptable errors (i.e., errors that could occur in any laboratory) from unacceptable errors that fall below the standard of care.²

Although the idea of blinded review has great merit, the methods for interpreting the results and their statistical validity should be verified. Currently, the statistics for blinded reviews are based on a non-reducible, false-negative fraction of 5%,³ a level of performance that to our knowledge has never been achieved.^{1,4,5} Exact thresholds for what constitutes acceptable, unacceptable, or uncertain errors should be defined. It would be helpful to incorporate the known performance level of the Pap test into the interpretation of the results and to measure the exact performance levels needed for blinded reviews. To address this issue, we sought to determine the power and statistical validity in an attempt to help improve multiple-slide blinded reviews (MSBR).

MATERIALS AND METHODS

For the purposes of this report, blinded review consists of review of a case in which all identifying information has been removed. Each case is screened the specified number of times by separate reviewers (i.e., independent reviews).

An important assumption of blinded review that was accepted for the purposes of the current study is that, provided the cytotechnologists who perform the review are trained appropriately and are certified to review cases, the results of blinded review that they achieve are an acceptable reflection of the skills provided by the cytology community. Certainly, different

cytotechnologists may possess different levels of skill, but these differences generally are not measurable. Therefore, it is assumed that, if the cytotechnologists performing the review are qualified appropriately, then the results of the review reflect the qualities of the case rather than the qualities of the cytotechnologists who perform the review.

The analysis performed in the current study was similar to that outlined previously for proficiency testing.^{6,7} Probabilities listed in Table 1 were calculated using the probability mass function of a binomial, random variable. Power and sample sizes were calculated using Nquery® software (Statistical Solutions, Saugus, MA)⁸ with the exact test for a single proportion using a 2-sided α value of 0.05 and $\geq 80\%$ power.

The results of MSBR reflect the likelihood that a case will be identified as abnormal. This is dependent on whether abnormal cells are present on the slide and how difficult these cells are to detect. For the purposes of this article, we defined “the ease of interpretation” of a case as the results demonstrated by blinded review. For example, a case that is diagnosed as abnormal 100% of the time has a relatively high ease of interpretation, higher than a case that is identified as abnormal only 70% or 80% of the time.

For comparison purposes, the ease of interpretation of specific types of cases can be estimated. Based on the results of large, blinded, cross-over studies, the ease of interpretation of a “routine” case can be estimated at 80%.¹ Although it is not defined as well and is dependent on how these cases are identified and measured, the ease of interpretation of cases that are missed on routine review (“routine false-negative cases”) has been estimated, based on repetitive blinded review, at 27%,⁹ 40%,¹⁰ and 43%.⁴ For the purposes of this report, we estimated the ease of in-

TABLE 2
Power to Detect a Difference between Ease of Interpretation of a Particular Case (Alternative Ease of Interpretation) and 80% with Two-Sided $\alpha = 0.05$ (Number of Reviews = 10)

Alternative ease of interpretation	Power (%)
0.10	99
0.20	96
0.30	84
0.40	63
0.50	37
0.60	16
0.70	4
0.90	1
0.95	1

terpretation screening of these “routine false-negative cases” as 40%.

RESULTS

Statistical Analysis

First, we sought to determine the theoretical probability that cases of a known ease of interpretation would give specific results using 10 blinded reviews.³ These results are shown in Table 1. For example, a case with an actual ease of interpretation of 80% (that is, if screened thousands of times by a typical cytology laboratory, it would be detected as abnormal 80% of the time) has a 68% chance of having a result ≥ 8 of 10 reviews and also has a $< 0.01\%$ chance of being detected as abnormal ≤ 3 times.

Second, we calculated the power to detect a difference in the proportion of results from 10 case reviews of an unknown case from that of a case with a known ease of interpretation of 80%. Table 2 shows that the power of the test (the ability of a test to identify accurately a significant difference that is present in the test population) depends strongly on the size of the difference that it is designed to detect. If a power of 80% (which is typical for this type of test) is required, then detecting a difference between 0% and 30% versus 80% is possible with only 10 reviews. Any smaller difference, however, would result in an underpowered test. The consequence of an underpowered test is that a statistically significant difference may not be identified by the test.

Third, we sought to determine how many reviews would be necessary to have 80% power to detect a difference between various hypothesized degrees of difficulty of the test case and various true or comparison degrees of difficulty. In Table 3, the top numbers are the sample sizes necessary to detect the specified difference in ease of interpretation (2-sided $\alpha = 0.05$; power = 0.80). The bottom numbers listed (in paren-

theses) are the lower and upper limits of abnormal results needed to demonstrate that the case's ease of interpretation is not significantly different from the hypothesized ease of interpretation. For example, if we wanted to show that a particular case had a hypothesized ease of interpretation of 90%, and if we wanted to show that it was statistically more likely to have this ease of interpretation than a true ease of interpretation of 80%, then we would need to obtain at least 96 abnormal reviews out of 113 reviews to support this. If ≤ 95 abnormal reviews are observed out of a total of 113 reviews, then we would conclude that the case's ease of interpretation is significantly less than the hypothesized 90%.

It is interesting to note that, as the true ease of interpretation approaches 100%, the number of reviews required to detect a set difference decreases. For example, detecting a 5% difference (90% vs. 95%) requires > 200 reviews, whereas detecting a 4% difference (95% vs. 99%) requires less. This is because of the increased true ease of interpretation, which allows for more certain results. The bigger the difference between the hypothesized ease of interpretation and the ease of interpretation it needs to be distinguished from, the fewer reviews will be required. It is possible (with 80% power) to detect a difference between a hypothesized ease of interpretation $\geq 90\%$ and a true ease of interpretation of 40% with only 7 reviews (Table 3). Attempting to distinguish a hypothesized ease of interpretation $\geq 40\%$ from a true ease of interpretation of 80% with only 10 reviews may result in $< 80\%$ power.

Determination of Clinically Relevant Thresholds

From the data in the literature, there appear to be at least 3 meaningful ease-of-interpretation levels that are worth evaluation as potential cutoff levels for interpreting this test: 99% (cases that almost always are identified), 80% (routine cases), and 40% (false-negative cases). From the data in Table 3, with a sample size of 20 (which would be used to distinguish between cases with a sensitivity of 99% and 80%), a case would have to be identified as abnormal 19 or 20 times to conclude that the ease of detection is not significantly different from 99%. An identification of abnormal ≤ 18 times would indicate that the ease of detection is significantly less than 99%. Although it is possible to calculate the results needed to distinguish between cases with an ease of detection of 40% and 99%, the data in the literature also strongly suggest that the vast majority of all cases have an ease of detection between these 2 thresholds, and these cases are not sorted adequately using this threshold. Finally, to show that a case is significantly more likely to have

TABLE 3
Sample Size to Detect Specified Differences in Ease of Interpretation (Two-sided $\alpha = 0.8$; Power = 0.80)

Hypothesized ease of interpretation	True ease of interpretation						
	40%	60%	80%	90%	95%	99%	99.9%
40%	NA	55 (16–28)	14 (3–9)	6 (0–4)	4 (0–3)	4 (0–3)	2 —
60%	60 (30–42)	NA	44 (21–32)	16 (6–13)	10 (3–8)	7 (2–6)	4 —
80%	15 (9–14)	50 (35–44)	NA	100 (73–86)	40 (28–36)	15 (9–14)	8 (4–8)
90%	7 (5–7)	19 (15–19)	113 (96–107)	NA	215 (186–200)	43 (35–44)	16 (12–16)
95%	7 (6–7)	15 (13–15)	45 (40–45)	250 (232–243)	NA	137 (126–134)	60 (54–60)
99%	5 (5)	8 (8)	20 (19–20)	54 (52–54)	150 (146–150)	NA	430 (422–429)
99.9%	5 (5)	8 (8)	17 (17)	36 (36)	72 (71–72)	555 (553–555)	NA —

NA: not applicable.

an ease of detection of 80% rather than an ease of detection of 40% requires that it be identified 9 of 15 times.

Alternatives to a Single Diagnostic Threshold

Determining the level of power in an analysis helps ensure that a case of a particular type is classified incorrectly no more than a particular percentage of the time. A commonly used threshold is 5% (that is, the number of reviews in a test will be sufficient so that a case is misclassified no more than 5% of the time). If the necessary number of reviews cannot be obtained, then alternative approaches can be used to ensure that at least the chance of a case being misclassified is no more than a specific percentage of the time, such as 5%. One approach is to split the reviews into two separate groups that are analyzed separately. This approach has been examined extensively for proficiency testing^{6,7} and will not be commented on further herein. Another alternative is to use more than one threshold. For example, a threshold may be set for cases that are identified regularly and reliably (e.g., 95%) that differs from the threshold set for cases that are not identified regularly and reliably (e.g., 50%) rather than using only a single threshold (typically, in between these 2 values) for both categories. The major consequence of such an approach is that it creates a group with indeterminate results that fall between the two values and that cannot be classified into one group versus another. Using this approach and the data in Table 1, the chance that a particular case will be misclassified can be reduced as low as possible by

changing the lower threshold; however, this does not increase the power of the test. The “cost” of limiting the chance of misclassification is taken in the increased number of cases that may be indeterminate and the decreased chance that a case will be classified correctly, factors that depend directly on which percentage the upper threshold is set.

Using this approach, the chance that a case will be classified correctly or incorrectly can be calculated using the current standards of 4 of 10 reviews and 6 of 10 reviews, respectively, for cases that are not regularly identified and are regularly identified. With these thresholds, the chance of a routine false-negative case (ease of detection, 40%) will be classified correctly as “not detected regularly and reliably” is 63%, and the chance that it will be misclassified as “regularly identified” is 16%. Conversely, the that chance a routine case (ease of detection, 80%) will be classified correctly as “regularly identified” is 96%, whereas the chance that a case will be misclassified as “not identified regularly” is 1%. Thus, using these criteria to sort routine cases, a routine case will be classified correctly 1.5 times more often than a routine false-negative case and will be misclassified 16 times less often.

These odds can be changed by changing the threshold used to define regularly identified and not regularly identified cases. If, instead of using thresholds of 4 of 10 reviews and 6 of 10 reviews, a threshold of 5 of 10 reviews was used for a case not identified regularly and a threshold of 7 of 10 reviews was used to identify a case reliably, then the chance of correctly classifying a routine case and a routine false-negative

case would be the same (84%), and the chance that it would be misclassified also would be the same (5%): Eleven percent of cases would be indeterminate.

DISCUSSION

The analysis of cervicovaginal specimens in cytology laboratories is not perfect and is associated with a known sensitivity between 70% and 80% under the best of circumstances, and this sensitivity applies to both ASC and low-grade squamous intraepithelial lesions.¹ Although nonblinded review of cervicovaginal cases in the context of litigation has been performed in the past, this largely has fallen into disfavor with the recognition of the relatively large routine error rate¹ and the increasing recognition of the potential for bias with this method.^{2,5,9} Instead, blinded review has become recognized increasingly as the most appropriate method with which to determine the standard of care.

The best method for interpreting the results of MSBR was investigated. In this regard, it was important to determine 1) what the interpretation is trying to show or, more specifically, the difference that the interpretation should demonstrate, and 2) what appropriate thresholds can be used in the future to measure this performance.

The first question is a value judgment, and there is no definite, single right answer. Instead, we have used the best available data concerning the known performance of the test to identify 3 reasonable thresholds to consider: cases that almost always are detected (ease of detection, 99%), routine cases (ease of detection, 80%), and routine false-negative cases (ease of detection, 40%). Certainly, other thresholds can be considered, but the use of other thresholds should require at least a rationale for choosing them.

These three thresholds can be used to identify three separate comparisons that may be useful with MSBR, and each of these combinations has their own particular advantages and disadvantages. Trying to distinguish between slides that are identified 99% of the time and 80% of the time has several advantages. First, at one level, it makes sense that the cases that virtually always are identified should be the cases that always are required to be identified. Second, it is important to distinguish these cases from the known average performance of most cases. However, to our knowledge this threshold is much higher than that used in blinded reviews to date, and it is unclear whether this threshold would be accepted in the setting of litigation. Trying to distinguish between cases that almost always are identified (ease of detection, 99%) and cases that routinely are missed (ease of detection, 40%) can be done but is of limited utility because the available evidence suggests that the ma-

ajority of cases have an ease of detection between those two values and would not be classified reliably by this measure.

Finally, trying to distinguish between cases that routinely are identified (ease of detection, 80%) and cases that routinely are missed (ease of detection, 40%) has the advantage of most closely resembling the current standard. However, this threshold, by its very definition, may be a difficult standard for the cytology community to meet because it requires cytotechnologists to be responsible for always identifying cases that, by definition, are not always identifiable. However, the fact that it is similar to the current standard suggests that this threshold at least should be considered. Therefore, for practical purposes, we believe the most obvious thresholds to consider in this process are the differences between cases with an ease of detection of 99% compared with 80%, and cases with an ease of detection of 80% compared with 40%.

According to the data provided in the current report, increasing the number of reviews can strengthen the power of MSBR. For comparisons of cases with an ease of interpretation of 99% and 80%, it may be necessary to review the case 20 times and, for cases with an ease of interpretation of 80% and 40%, it may necessary to review the case 15 times to achieve a standard level of statistical power. Of course, it may not always be necessary to review each case all 15 or 20 times. For example, for the threshold of 99% and 80%, as soon as a case is missed twice, there is no need to review it further; it already has shown that it is significantly more likely to represent a case with an ease of detection of 80%.

However, blinded review is a labor-intensive process, and there may be advantages to reviewing cases only 10 times. In this setting, one may not be able to achieve standard statistical power to make these distinctions. However, by using two separate thresholds, one can at least limit the chance of misclassifying a case. Comparison of cases with an ease of detection of 40% and 80% using thresholds of 5 of 10 reviews and 7 of 10 reviews, respectively, may strengthen MSBR. In this setting it may be stated confidently that the chance of misclassifying a case with an ease of detection of 40% is identical to the chance of misclassifying a case with an ease of detection of 80% (approximately 5%).

In the current study, we provided data that can be used easily to interpret the results of blinded review in a statistically meaningful manner. The data suggest that, for the most clinically relevant distinctions, it may be necessary to increase the number of reviews to 15 or 20 (depending on the threshold chosen) or to use cutoff levels of 5 of 10 reviews and 7 of 10 reviews.

Nevertheless, it is unlikely that every case will need to be reviewed all 15 or 20 times to obtain a statistically significant result.

REFERENCES

1. Renshaw AA. Measuring sensitivity in gynecologic cytology. A review. *Cancer (Cancer Cytopathol)*. 2002;96:210–217.
2. Fitzgibbons PL, Austin RM. Expert review of histologic cases and Papanicolaou tests in the context of litigation or potential litigation. Surgical Pathology Committee and Cytopathology Committee of the College of American Pathologists. *Arch Pathol Lab Med*. 2000;124:1717–1719.
3. Holladay EB, Austin RM. Multiple case blinded rescreening for review of litigation cases. In: Allen KA, Holladay EB, editors. Risk management for the cytology laboratory. Raleigh, NC: American Society of Cytotechnologists, 2001:149–162.
4. Renshaw AA. Estimating the percentage of Papanicolaou smears that can be reproducibly identified: modeling Papanicolaou smear interpretation based on multiple blinded rescreenings. *Cancer (Cancer Cytopathol)*. 2001;93:241–245.
5. Renshaw AA. Experts in Wonderland: in search of the right test and the scientific method. *Diagn Cytopathol*. 2000;23:297–298.
6. Nagy GK, Collins DN. False-positive and false-negative proficiency test results in cytology. *Acta Cytol*. 1991;35:3–7.
7. O’Leary TJ. Proficiency testing in cytopathology. *Accredit Qual Assur*. 2002;7:357–361.
8. Statistical Solutions. NQuery Advisor 4.0 software. Saugus, MA: Statistical Solutions, 1995.
9. Renshaw AA, Lezon KM, Wilbur DC. The human false-negative rate of rescreening Pap tests. Measured in a two-arm prospective clinical trial. *Cancer (Cancer Cytopathol)*. 2001;93:106–110.
10. O’Sullivan JP, Chapman PA, Jenkins L, Smith R. Characteristics of high grade dyskaryotic cervical smears likely to be missed on rapid rescreening. *Acta Cytol*. 2000;44:37–40.