Discordance in the Histopathologic Diagnosis of Melanoma and Melanocytic Nevi Between Expert Pathologists

EVAN R. FARMER, MD, RENÉ GONIN, PHD, AND MARK P. HANNA, MS

The reliability of a diagnostic test depends on the reproducibility of the result. Many clinical diagnostic tests can be quantified with established ranges and standard deviations. Other tests are more subjective, such as those that depend on analysis of a visual image with an increased possibility of variance in the result. To study this variance, the authors analyzed the performance of expert pathologists in the interpretation of cutaneous melanocytic tumors. A panel of expert pathologists was convened to review anatomic pathology specimens from melanocytic tumors. Each pathologist submitted five specimens, from which 37 were selected for review. Only one slide was used for each case. All specimens were interpreted by each pathologist without consultation with each other. In addition to standard diagnostic terms, each specimen was designated as benign, malignant, or indeterminate. Statistical analysis was used to determine the degree of concordance. The combined κ statistic for the eight observers and three possible outcomes (benign, malignant, or indeterminate) was 0.50. A κ statistic of this magnitude is defined as being moderate. In 62% of the specimens, there was unanimous agreement or only one discordant designation. Thirty-eight percent had two or more discordant interpretations. No single pathologist had a disproportionate number of discordant designations. This study mimics the consultation practice of anatomic pathology and shows the variability and discordance in diagnostic language and designation of biological behavior. The results suggest the criteria for the diagnosis of melanomas and melanocytic nevi need to be refined and more consistently applied. HUM PATHOL 27:528-531. Copyright © 1996 by W.B. Saunders Company

Key words: **observer variation, reproducibility of results, melanoma, nevus, Cohen's κ, pathology, human.**

The efficacy of a diagnostic test can be defined as the ability of the test to indicate the presence or absence of a disease, or to make a prediction about the future behavior of the disease. The reliability of a diagnostic test first depends on the reproducibility of the interpretation or result on a given specimen. Many tests can be quantified with the results expressed as numbers or units with established ranges and standard deviations for a given population. Other tests are more subjective, such as those that depend on analysis of a visual image. Routine radiological tests and anatomic pathology tests are examples of tests that depend on an observer's analysis of a visual image. Because these tests depend on subjective interpretation, there exists the possibility of variance in the reported result. It is a common experience in clinical medicine to have various opinions on a single radiological or anatomic pathology test. One possible solution to decrease this variance in interpretation is to have panels of experts review the more serious diseases and offer a diagnosis by consensus.[1-3] Although this approach has merit, it is not yet practical, and many rely on individual experts for consultation. One important question is: How well do the experts agree on a given case when they interpret it independently?

As part of the planning for the National Institutes of Health Consensus Development Conference on the Diagnosis and Treatment of Early Melanoma,[4] a panel of expert pathologists and dermatopathologists in melanoma diagnosis was convened to review anatomic pathology specimens from melanomas and melanocytic nevi. The intent of this review was to develop definitions of histopathologic features and criteria for diagnosis that could be used in the subsequent conference as a way of standardizing the data.[5] Panel members reviewed the specimens in their own laboratory in the practice of consultation cases, signing them out in their individual routine fashion. As there were no prior meetings of the panel members to discuss criteria for this conference, and the panel members did not know each others' diagnoses, this strategy would mimic the practice of consultation pathology regarding melanomas and melanocytic nevi. This article presents an analysis of their interpretations.

METHODS

The expert panel of pathologists was selected on the basis of their reputation and publications in the diagnosis of melanomas and melanocytic nevi. They were from different medical centers and held various viewpoints on the criteria for diagnosis. The panel members did not know the results of each others' diagnoses, by mutual agreement, for the consultation phase of this project. Anonymity was discontinued once all of the consultations were completed.

Each panel member was requested to submit five cases of melanomas or melanocytic nevi that shared histological features with melanomas. These cases were to be considered "classic cases" by the submitter. The cases could be unusual or difficult, but had to be representative of a case that the submitter could use for publication as an example of a specific entity. Only one slide from each case was used. The original history from each patient was provided, but the clinical diagnosis was deleted.

From the cases submitted by the expert panel, 35 cases were selected by the planning committee to be reviewed. Two additional cases were added by the planning committee to represent entities that had not been included by the panel

From the Departments of Dermatology and Medicine and the Section of Biostatistics, Indiana University School of Medicine, Indianapolis, IN. Accepted for publication November 21, 1995.

Address correspondence and reprint requests to Evan R. Farmer, MD, Department of Dermatology, Indiana University School of Medicine, 550 N. University Blvd, Suite 3240, Indianapolis, IN 46202.

0046-8177/96/2706-0005$5.00/0

TABLE 1. Examples of Diagnostic Language

Observer	Case No. 13	Case No. 16	Case No. 21
1	Malignant melanoma, invasive, Clark's level III	Malignant melanoma in situ	Malignant melanoma, Clark's level II
2	Malignant melanoma, superficial spreading, Clark's level III	Junctional lentiginous nevus with atypia	Compound dysplastic nevus, pigmented spindle cell
3	Malignant melanoma, nodular, Clark's level III	Compound lentiginous dysplastic nevus with atypia	Compound dysplastic nevus with atypia
4	Malignant melanoma, nodular, Clark's level IV	Junctional nevus with atypia	Malignant melanoma, superficial spreading, Clark's level III
5	Compound nevus, Spitz's nevus	Malignant melanoma in situ	Malignant melanoma
6	Spindle and epithelioid cell nevus, inflamed	Lentigo maligna	Malignant melanoma, superficial spreading
7	Epithelioid cell nevus, atypical	Atypical melanocytic hyperplasia	Compound dysplastic nevus
8	Malignant melanoma	Malignant melanoma in situ	Malignant melanoma

members. All identifying marks were removed from the slides, and they were relabeled with a standard label for the project. The cases were then randomized to prevent clustering by diagnosis or submitter. Once randomized, the original diagnoses and submitters were unknown to the planning committee to minimize bias in analysis. All 37 cases, along with the original history, were sent to each panel member. The panel member then sent the cases back to the planning committee before being sent to the next panel member. Thus, all panel members based their diagnoses on study of the same glass slide. All panel members were requested to sign out each case in their usual fashion. After the sign-out was completed, they were requested to define the terms they used and to provide the criteria used to establish the particular diagnosis. Each case was also to be designated as benign, malignant, or indeterminate. Indeterminate was defined as unable to provide a definitive diagnosis of either benign or malignant, and did not imply a biological position or behavior between the two poles. The designations benign, malignant, and indeterminate were considered to be independent variables.

Statistical Analysis

To measure the rate of agreement among expert pathologists, the generalized κ statistic of Landis and Koch[6] was used. This approach is essentially an analysis of variance approach because both the number of raters and rating categories were greater than two. The 95% confidence interval for the true κ was calculated using the asymptotic standard error for κ.[7] All calculations were performed using the Stata package (Stata Corporation, College Station, TX).[8]

RESULTS

All 37 cases were reviewed by each panel member, and the diagnoses, definitions, criteria, and designations of benign, malignant, or indeterminate were completed. There was no breakage of slides, and no slides were lost or replaced.

As expected, there was variability in the terms used for diagnosis (Table 1). The prognostic designations of benign, malignant, or indeterminate, as independent variables, proved to be better categories for comparison of concordance than the actual diagnoses. Thirteen cases had unanimous agreement: eight benign and five malignant. The 24 cases that lacked unanimous concordance are displayed in Table 2; all 37 are summarized by slide in Table 3 and by observer designation in Table 4. The clusters of discordant designations are presented in Table 5. There were 62% of the cases in which there was unanimous agreement or only one discordant designation. Thirty-eight percent had two or more discordant interpretations. No single pathologist had a disproportionate number of discordant designations (Table 2).

The combined κ statistic for the eight observers and three possible outcomes (benign, malignant, or indeterminate) was 0.50. This was found to be significantly different from none (Pvalue $< .0001$) using the appropriate test statistic.[7] However, by definition, a κ statistic of this magnitude is considered to indicate only moderate agreement.[9]

DISCUSSION

Although this study originated as part of the planning process for the National Institutes of Health Consensus Development Conference on the Diagnosis and Treatment of Early Melanoma,[4] it mimics the actual, consultation practice of anatomic pathology. The cases were selected for their "classic features," and then were interpreted by a panel of eight expert pathologists in the diagnosis of melanoma and melanocytic nevi. The panel members themselves each contributed five cases to the study, from which the total set of 37 cases was chosen. The panel was initially anonymous to each other and did not participate in any meetings before interpreting the cases for the purposes of this study. Each panel member interpreted the cases in their own laboratories using their own standard language and criteria for diagnosis. They then designated each case as benign, malignant, or indeterminate (the latter designation meaning that the case did not clearly fit their criteria for being either benign or malignant).

This study shows the variability and discordance in diagnostic language and designation of biological behavior (benign, malignant, or indeterminate) that is possible on a given case of a melanocytic tumor. The κ statistic of 0.50 indicates only moderate agreement between the observers on these cases. By definition, a κ statistic of 0.81 or greater reflects excellent to almost

TABLE 2. Cases Lacking Complete Concordance

	Case Nos.																							
Observer	1	2	6	7	8	9	13	14	15	16	17	19	21	22	25	26	27	28	29	30	33	34	35	37
1	I	B	M	B	I	M	M	B	B	M	B	M	M	B	B	M	M	M	M	B	B	M	M	M
2	B	B	B	M	I	B	M	B	B	B	B	M	B	M	B	I	M	M	I	B	B	M	M	B
3	B	B	B	M	B	B	M	B	M	B	B	M	B	M	B	M	M	M	M	B	B	M	M	I
4	B	M	B	M	B	B	M	B	B	B	B	M	M	B	M	M	M	B	M	B	B	M	M	M
5	M	B	B	M	B	B	B	B	B	M	B	M	M	B	B	M	M	M	M	M	B	M	M	M
6	M	B	M	M	B	M	B	M	B	M	B	M	M	M	B	M	M	M	M	M	B	M	M	B
7	B	B	B	M	B	B	B	B	B	B	B	B	B	B	B	M	M	M	B	B	B	M	B	I
8	B	M	M	M	B	B	M	M	B	M	M	M	M	B	M	M	I	M	I	M	I	I	M	B

Abbreviations: B, benign; I, indeterminate; M, malignant.

perfect agreement, values between 0.61 to 0.80 represent substantial agreement, values between 0.41 to 0.60 represent moderate agreement, values between 0.21 to 0.40 are indicative of fair agreement, and values between 0.01 to 0.20 are indicative of slight agreement, meaning slightly above that achieved by chance alone.[9] Complete agreement or only one discordant interpretation was achieved in only 62% of the 37 cases. Thirty-eight percent had two or more discordant interpretations. Although the cases in this study were selected and interpreted by an expert panel, it could be anticipated that unselected cases interpreted by pathologists, dermatopathologists, and dermatologists may show an even greater degree of discordance.

Prior studies of concordance in the diagnosis of melanomas and dysplastic nevi using κ statistics have noted similar findings of only fair to moderate agreement in the recognition of presence, absence, or degree of individual histological criteria.[10-17] These studies and this one suggest that although histological criteria can be described and published, the consistent application of these criteria by various observers has not been achieved. This is probably caused by the subjective nature of many histological criteria. Criteria, such as tumor thickness or the presence or absence of ulceration, seem to be more consistently applied.[11,14] Almost all of these published studies differed from the current one by having pretest meetings to agree on criteria, having the cases from the same institution as the observers, or by having the observers from the same institution or training program.[10-16] One recent study of dysplasia in melanocytic nevi used six observers from different institutions without previously agreed-on criteria and found an overall fair concordance with a κ of 0.34.[17]

Similar problems with interobserver concordance have been noted in the histological diagnosis of cutaneous T-cell lymphoma.[1] In that study, the panel as a group provided more consistency in diagnosis than did individual observers. The use of panels for decreasing histological variability in the evaluation of cases has been previously suggested in other studies.[2,3] Radiologists also show significant variability in interpretation of images. In one recent study of interpretations of mammograms, the weighted κ was 0.47 between pairs

TABLE 3. Summary of Status by Slide

Slide No.	Malignant	Benign	Indeterminant
1	2	5	1
2	2	6	0
3	0	8	0
4	8	0	0
5	0	8	0
6	3	5	0
7	7	1	0
8	0	6	2
9	2	6	0
10	0	8	0
11	8	0	0
12	0	8	0
13	5	3	0
14	2	6	0
15	1	7	0
16	4	4	0
17	1	7	0
18	8	0	0
19	7	1	0
20	8	0	0
21	5	3	0
22	3	5	0
23	0	8	0
24	8	0	0
25	2	6	0
26	7	0	1
27	7	0	1
28	7	1	0
29	5	1	2
30	3	5	0
31	0	8	0
32	0	8	0
33	0	7	1
34	7	0	1
35	7	1	0
36	0	8	0
37	3	3	2

TABLE 4. Observer's Designation

Observer	Malignant	Benign	Indeterminate
1	18	17	2
2	13	21	3
3	16	20	1
4	17	20	0
5	18	19	0
6	21	16	0
7	10	26	1
8	19	14	4

TABLE 5. Clustered Concordance on All 37 Cases

Concordance	Cases (%)
Complete agreement	13 (35)
One discordant	10 (27)
Two or more discordant	14 (38)

of radiologists.[18] There was also a substantial disagreement in management recommendations by the radiologists based on their interpretations.

Interobserver concordance is also problematic in the clinical diagnosis of melanocytic tumors or recognition of clinical features.[19-n22] The clinical diagnostic accuracy of melanomas for university-based physicians in one oncology center was 64%.[19] Numerically derived criteria also pose a problem in concordance. In one study of nevus counts on one arm, there was only a moderate concordance between two dermatologists with a κ of 0.51. Nonphysician interviewers compared with the dermatologists had a poor concordance, with a κ of 0.19.[22]

In this study, the authors did not examine intraobserver concordance on either the presenter's own cases or on the set of 37 as a whole. Because each panel member contributed only five cases and there was only one slide from each one, it seems reasonable there would have been a high degree of concordance as pathologists tend to remember previously read slides. In one study of histological dysplasia, intraobserver agreement was substantial, with an overall average κ value of 0.63.[17]

By study design, the authors did not know the exact diagnosis of each case as determined by patient outcome. The intent of the study was to examine the concordance of diagnostic interpretation as it would occur in the practice of anatomic pathology consultation. Patient outcomes might not have been available on all of the cases for many years.

The authors did not evaluate the individual histological criteria used in these cases for concordance as there was a large variation in language and meaning. The published studies that have assessed the individual histological variables have relied on pretest meetings and agreed-on definitions and criteria.[11,16] Prior agreement on definitions and criteria is required to ensure the standardization of the measurements.

In summary, concordance in the histological diagnosis of melanomas and melanocytic nevi is moderate at best, even by experts in melanocytic tumors. Until criteria for the diagnosis of melanomas and melanocytic nevi can be refined and more accurately applied, this discordance probably will not improve. The use of panels of pathologists to study individual specimens may decrease this histological variability but will need further study to verify efficacy to warrant its widespread use.

Acknowledgment. The authors are indebted to Drs A. Bernard Ackerman (Jefferson Medical College); Wallace H. Clark, Jr (Kennebunk, ME); Harry L. Evans (M.D. Anderson Medical Center); Robert G. Freeman (Dallas, TX); James H. Graham (Del Mar, CA); Martin H. Mihm, Jr (Albany Medical College); Juan Rosai (Memorial Sloan-Kettering Cancer Center); and Richard W. Sagebiel (University of California, San Francisco–Mount Zion Cancer Center) for serving as the panel of expert pathologists; and to Drs Alan N. Moshell and Stephen I. Katz (National Institutes of Health), and Dr Lowell A. Goldsmith (University of Rochester School of Medicine) for their organization of the "Workshop Without Walls," which served as the basis for this study.

REFERENCES

1. Olerud JE, Kulin PA, Chew DE, et al: Cutaneous T-cell lymphoma: Evaluation of pretreatment skin biopsy specimens by a panel of pathologists. Arch Dermatol 128:501-507, 1992

2. Prade M, Sancho-Garnier H, Cesarini JP, et al: Difficulties encountered in the application of Clark classification and the Breslow thickness measurement in cutaneous malignant melanoma. Int J Cancer 26:159-163, 1980

3. Suffin SC, Waisman J, Clark WH Jr, et al: Comparison of the classification by microscopic level (stage) of malignant melanoma by three independent groups of pathologists. Cancer 40:3112-3114, 1977

4. Anonymous: Diagnosis and treatment of early melanoma: NIH Consensus Development Conference, January 27-29, 1992. Consensus Statement 10:1-25, 1992

5. Clark WH Jr, Evans HL, Everett MA, et al: Early melanoma: Histologic terms. Am J Dermatopathol 13:579-582, 1991

6. Landis JR, Koch GG: A one-way components of variance model for categorical data. Biometrics 33:671-679, 1977

7. Fleiss JL, Nee JCM, Landis JR: Large sample variance of kappa in the case of different sets of raters. Psychol Bull 86:974-977, 1979

8. Stata Corporation: Stata Reference Manual: Release 3.1. (ed 6). (Publication no. 2, 420-428). College Station, TX, Stata Corporation, 1993

9. Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 33:159-174, 1977

10. Holman CD, James IR, Heenan PJ, et al: An improved method of analysis of observer variation between pathologists. Histopathology 6:581-589, 1982

11. Heenan PJ, Matz LR, Blackwell JB, et al: Inter-observer variation between pathologists in the classification of cutaneous malignant melanoma in western Australia. Histopathology 8:717-729, 1984

12. Ahmed I, Piepkorn MW, Rabkin MS, et al: Histopathologic characteristics of dysplastic nevi: Limited association of conventional histologic criteria with melanoma risk group. J Am Acad Dermatol 22:727-733, 1990

13. Duray PH, DerSimonian R, Barnhill R, et al: Inter-observer recognition of dysplastic nevi from a mixed group. J Cutan Pathol 17:290, 1990 (abstr)

14. Colloby PS, West KP, Fletcher A: Observer variation in the measurement of Breslow depth and Clark's level in thin cutaneous malignant melanoma [see comments]. J Pathol 163:245-250, 1991

15. Duncan LM, Berwick M, Bruijn JA, et al: Histopathologic recognition and grading of dysplastic melanocytic nevi: An interobserver agreement study. J Invest Dermatol 100:318S-321S, 1993 (suppl)

16. Hastrup N, Clemmensen OJ, Spaun E, et al: Dysplastic naevus: Histological criteria and their inter-observer reproducibility. Histopathology 24:503-509, 1994

17. Piepkorn MW, Barnhill RL, Cannon-Albright LA, et al: A multiobserver, population-based analysis of histologic dysplasia in melanocytic nevi. J Am Acad Dermatol 30:707-714, 1994

18. Elmore JG, Wells CK, Lee CH, et al: Variability in radiologists' interpretations of mammograms. N Engl J Med 331:1493-1499, 1994

19. Kopf AW, Mintzis M, Bart RS: Diagnostic accuracy in malignant melanoma. Arch Dermatol 111:1291-1292, 1975

20. Barnhill RL, Roush GC, Ernstoff MS, et al: Interclinician agreement on the recognition of selected gross morphologic features of pigmented lesions: Studies of melanocytic nevi V. J Am Acad Dermatol 26:185-190, 1992

21. Miller M, Ackerman AB: How accurate are dermatologists in the diagnosis of melanoma? Degree of accuracy and implications. Arch Dermatol 128:559-560, 1992

22. Byles JE, Hennrikus D, Sanson-Fisher R, et al: Reliability of naevus counts in identifying individuals at high risk of malignant melanoma. Br J Dermatol 130:51-56, 1994