Published in final edited form as:

J Cutan Pathol. 2018 July ; 45(7): 478–490. doi:10.1111/cup.13147.

Complexities of Perceived and Actual Performance in Pathology Interpretation: A Comparison of Cutaneous Melanocytic Skin and Breast Interpretations

Patricia A. Carney, PhD¹, Paul D. Frederick, MPH, MBA², Lisa M. Reisch, PhD², Linda Titus, PhD, MA³, Stevan R. Knezevich, MD⁴, Martin A. Weinstock, MD⁵, Michael W. Piepkorn, MD⁶, Raymond L. Barnhill, MD⁷, David E. Elder, MB, ChB, FRCPA⁸, Donald L. Weaver, MD⁹, and Joann G. Elmore, MD, MPH¹⁰

¹Professor of Family Medicine, Oregon Health & Science University, Portland, OR

²Department of Internal Medicine, University of Washington School of Medicine, Seattle, WA

³Departments of Epidemiology and of Pediatrics, Geisel School of Medicine at Dartmouth, and the Norris Cotton Cancer Center, Lebanon, NH

⁴Pathology Associates, Clovis, CA, USA

⁵Professor of Dermatology, The Warren Alpert Medical School of Brown University, Providence, RI

⁶Division of Dermatology, Department of Medicine, University of Washington School of Medicine, Seattle, WA; Dermatopathology Northwest, Bellevue, WA

⁷Department of Pathology, Institut Curie, University of Paris Descartes, Paris, France

⁸Department of Pathology, University of Pennsylvania, Philadelphia, PA

⁹Professor of Pathology, University of Vermont, Burlington, VT

¹⁰Professor of Internal Medicine, University of Washington, Seattle, WA

Abstract

Background—Little is known about how pathologists process differences between actual and perceived interpretations.

Objective—To compare perceived and actual diagnostic agreement before and after educational interventions.

Methods—Pathologists interpreted test sets of skin and/or breast specimens that included benign, atypical, in situ and invasive lesions. Interventions involved self-directed learning, one skin and one breast, that showed pathologists how their interpretations compared to a reference diagnoses. Prior to the educational intervention, participants estimated how their interpretations would compare to the reference diagnoses. After the intervention, participants estimated their overall agreement with the reference diagnoses. Perceived and actual agreement were compared.

Results—For pathologists interpreting skin, mean actual agreement was 52.4% and overall pre- and post-interventional mean perceived agreement was 72.9% vs. 54.2%, an overestimated mean difference of 20.5% (95% confidence interval 17.2% to 24.0%) and 1.8% (95% CI –0.5% to 4.1%), respectively. For pathologists interpreting breast, mean actual agreement was 75.9% and overall pre- and post-interventional mean perceived agreement was 81.4% vs 76.9%, an overestimation of 5.5% (95% CI 3.0% to 8.0%) and 1.0% (95% CI 0.0% to 2.0%), respectively.

Conclusions—Pathologists interpreting breast tissue had improved comprehension of their performance after the intervention compared to pathologists interpreting skin lesions.

Keywords

continuing medical education; breast pathology; dermatopathology

Introduction

Diagnostic complexities exist across the spectra of histopathology with recent studies of this phenomenon focusing on breast and cutaneous lesions (1–6). Collectively, these studies indicate that for both breast and cutaneous lesions, distinguishing between atypia and in situ disease is especially challenging (1–6). Though these studies indicate these intricacies exist, less discussed is that misclassifications can result in both under- and over- diagnosis, affecting subsequent treatment regimens across the entire diagnostic range (7, 8), and more information is needed about how interventions might be developed to address such complexities. Most pathologists who are uncertain of a complex diagnosis will obtain a second opinion (9), and now more than 60% of pathology laboratories in the U.S. have policies to guide the use of second opinions (10). In a recent cross-sectional study involving pathologists (11), survey respondents reported that the use of second opinions improves their own diagnostic accuracy. This suggests the learning curve for expertise development expands beyond residency and into independent practice.

Some human malignancies have similar morphologic traits (12), a phenomenon that is fairly common in surgical pathology. Known as morphologic mimicry, one such cluster includes ductal carcinomas of cutaneous appendages, breasts, and salivary glands (12). In one study of 103 tumors in this structural cluster, striking homologies among these tumors were found that were typified by irregular permeative clusters and cords of atypical polygonal cells with varying luminal differentiation (12). While most of the literature on diagnostic agreement of histopathology includes studies within one tissue type, we hypothesized that much could be learned by studying levels of diagnostic agreement across different tissue types when interpretations were compared with a reference diagnosis obtained by a panel of experienced pathologists with expertise in their respective tissues.

We conducted an observational study that involved enrolled pathologists who interpret skin and breast biopsies, all of whom interpreted sets of cases representing a relevant spectrum of diagnostic categories in their respective fields. More specifically, we designed two web-hosted self-paced continuing medical education interventions, one focused on melanocytic skin biopsy interpretations and the other focused on breast biopsy interpretations of ductal epithelial proliferations. Each intervention allowed participants to compare their

interpretations to a reference diagnosis developed by a consensus panel. Prior to seeing their how their interpretations compared to the reference diagnoses but after interpreting the test sets, we asked participants to estimate the extent to which their interpretations would agree with the reference diagnosis so we could explore differences between perceived and actual performance pre-intervention. After the intervention, we asked pathologists to recall how their interpretations compared to the reference diagnosis to determine differences between perceived and actual performance post-intervention. Here, we present the results from these interventions and discuss how findings from this study may help unite pathologists toward coming to a common understanding of interpretations of difficult or borderline cases.

Methods

IRB, Physician Recruitment and Survey

Institutional review board (IRB) approval was obtained from the University of Washington/Fred Hutchinson Cancer Research Center, Dartmouth College, the University of Vermont, and Providence Health & Services of Oregon.

All participants were enrolled in one of two nation-wide studies on diagnostic variability, one focused on melanocytic cutaneous lesions (13) and the other focused on interpretation of breast biopsy interpretations (3). Pathologists who interpret skin biopsies were recruited using publicly available information from ten U.S. states (CA, CT, HI, IA, KY, LA, NJ, NM, UT, and WA). Pathologists who interpret breast biopsies were recruited using the same approach from eight geographically diverse states (AK, ME, MN, NM, NH, OR, WA, VT). More specifically, eligible pathologists were invited to participate via email, with telephone and street mail used for follow-up of initial non-responders. Eligibility criteria included interpreting cutaneous melanocytic or breast biopsies as part of current practice, having signed out biopsies for at least one year prior to enrollment, and intention to continue interpreting biopsies for at least one year after enrollment.

All participating physicians completed a brief 10-minute survey that assessed their demographic characteristics (age, sex), training and clinical experiences (fellowship, case load, interpretive volume, years interpreting, academic affiliations), and perceptions of how challenging their respective pathology types are to interpret. This survey was administered prior to beginning test set interpretations. Two hundred and seven pathologists were enrolled in the skin biopsy specimen study (Figure 1), and of these 187 completed all pre-educational intervention study activities and were therefore eligible to participate in the educational intervention (90.3%), of whom 158 completed it (84.5%). A full description of the overall pool of pathologists who interpreted skin lesions is described elsewhere (13), but those participating in the intervention portion of the study represented 52.5% of pathologists (158/301) whose eligibility we could determine from the ten states.

One hundred and twenty-six pathologists who interpret breast biopsy specimens agreed to take part in the study and were eligible to undertake the educational intervention according to completion of all pre-intervention study activities. Of these, 94 completed the intervention (74.6%), with 92 providing complete information (73.0%) (Figure 1). A full description of the overall pool of pathologists who interpreted breast tissue is described elsewhere (3), but

those participating in the intervention portion of the study represented 37.3% of pathologists (94/252) whose eligibility we could determine from the eight states. Four pathologists who completed the intervention portion of the breast interpretation study also participated in the skin pathology study, which occurred approximately two years later.

Test Set Development

Detailed information on the development of the test sets is published elsewhere for cutaneous melanocytic lesions (5) and breast specimens (14). Briefly, cutaneous test set cases were identified from a large dermatopathology laboratory in the Pacific Northwest (5), and breast test set cases were identified from biopsy specimens obtained from mammography registries with linkages to breast pathology and/or tumor registries in Vermont and New Hampshire (15, 16).

A consensus process was undertaken by three experienced dermatopathologists and three experienced breast pathologists to come to a consensus reference diagnosis for each case in the test sets, which is described in detail elsewhere (5, 14). For cutaneous melanocytic cases, 240 unique cases were randomly assigned into one of five test sets, which contained 48 cases each and represented the following five diagnostic categories: 1) nevus/mild atypia (10.4%), 2) moderate atypia (15%), 3) severe atypia/melanoma in situ (25%), 4) pT1a invasive melanoma (24.2%), and 5) pT1b or higher stage invasive melanoma (25.4%). For breast pathology cases, 240 unique patient cases were randomly assigned into one of four test sets, which contained 60 cases each and represented the following four diagnostic categories: 1) benign without atypia (30%), 2) atypia (30%), 3) ductal carcinoma in situ (30%) and 4) invasive cancer (10%). These distributions resulted in oversampling more complex interpretive cases, which would help us quantify the diagnostic challenges to be included in an educational intervention we designed for participants and delivered at the end of the study.

Educational Intervention Development and Implementation

The educational intervention was designed to provide a research-based review of differences among pathologists in cutaneous melanocytic and breast lesion interpretation. It was a self-paced Internet hosted program individualized to participant's performance on their respective test sets. Figure 2 illustrates activities related to the educational intervention.

Pre-intervention Assessments—At the start of the program (before participants reviewed how their interpretations compared to the reference diagnosis), we asked them to compare how similar the test set cases were to cases they see in actual clinical practice (response options ranged from 'I never see cases like these' to 'I always see cases like these'), and we asked them to indicate the number of continuing medical education (CME) hours they had undertaken in cutaneous melanocytic or breast or lesions as well as their CME preferences (instructor led, self-directed or other). Additionally, we asked participants to estimate the proportion of their diagnoses on test cases they thought would agree with the consensus reference diagnoses overall and within each of the four or five diagnostic categories reviewed. These estimates are referred to as the "pre-intervention perceived agreement".

Operational Features of the Intervention—The educational intervention then proceeded to show the pathologists, on a case-by-case basis, how their independent interpretations actually compared with the consensus reference diagnosis according to diagnostic classification. By clicking on the case number, participants could view a digital whole slide image of the case on their computer screens and could scan and magnify the image as with a conventional microscope. Case-specific microscopic diagnostic comments, developed by experienced pathologists on the respective consensus reference panels, were provided for each case. Participants could review the cases and comments for as long as they desired and were given the opportunity to share their own thoughts on each case after reviewing their results and reading the experienced pathologists' microscopic diagnostic comments.

Post-Intervention Assessments—After the intervention, a survey asked study participants to re-rate how their interpretations compared to the consensus reference diagnosis, so we could assess the extent to which they recognized, processed and recalled the difference between their perceived and actual agreement with the consensus panel as a result of the intervention. These estimates are referred to as “post-intervention perceived agreement”. We then asked them to complete required knowledge questions, so we could award CME credits. Completion of all study activities, including the test set interpretations and the educational intervention resulted in awarding up to 20 Category 1 CME hours.

Data Analyses

We used descriptive statistics to characterize the demographic and clinical training experience as well as ratings of the test sets and challenges involved in interpreting pathology. We stratified the pathologists who interpreted skin tissue according to their receipt of specialized training and/or board certification in dermatopathology. Only three of the pathologists interpreting breast biopsies were fellowship trained in breast pathology, a number too small for stratification. Histograms were used to compare perceived differences in interpretive performance compared to actual performance, both overall as well as within each diagnostic category for both pre- and post-intervention assessments. Among the 94 participants who completed the breast CME program, differences between perceived and actual performance rates could not be assigned for 2 participants (2.1%) due to missing responses. These two participants were excluded from analyses. Similarly, 4 (6.0%) of the 67 enrolled participants who completed the dermatopathology CME program had incomplete or missing responses and were excluded from analyses (Figure 1). Categorical data are presented as frequencies and percentages and for continuous variables, values are reported as the mean and standard deviation (SD).

Pathologists' estimates for each of the disease category were assigned a weighted average according to the proportion of their diagnoses they estimated would agree with the reference' diagnoses and, as the weighting pattern, the number of cases each participant interpreted within each diagnostic class. We subtracted the weighted estimate of agreement with expert consensus diagnoses from the participants' actual agreement prior to and after the intervention, separately, to yield the difference between perceived performance and actual performance.

To assess normality of the distributions of performance measures, we graphically inspected frequency distributions and box plots of the differences between perceived and actual performance. We performed a parametric paired-samples t-test to test the hypothesis that a difference between perceived and actual performance rates exists in either the pre- or post-intervention assessments according to each diagnostic category and overall performance according to each educational intervention program. Mean differences were computed as the subtracted difference between the mean value of perceived and actual agreement with 95% confidence intervals (CI) also reported. A positive mean difference indicates an overestimation of performance and a negative difference indicates an underestimation of performance relative to the participants' actual performance. We defined alignment as the proportion of participants whose difference in perceived versus actual agreement fell within $\pm 5\%$ of zero difference.

Categorical data were compared by chi-square test (or Fisher's exact test when appropriate) and for continuous data, by one-way ANOVA on either ranked (Kruskal-Wallis) or original (unranked) data. We examined three pairwise differences of continuous data and proportions across the four groups representing each of two pathology specialties (skin pathologists and breast pathologists) and the two categories comprising all skin pathologists consisting of dermatopathology certified and/or fellowship trained and skin pathologists having no specialty training in dermatopathology. To avoid chance findings, a conservative Bonferroni correction was applied using an adjusted alpha level of 0.017, which is equal to the original alpha level divided by the number of tests (0.05/3).

To test whether the mean difference between perceived and actual performance was influenced by dermatopathology specialty, we regressed perceived performance (a continuous dependent variable) on dermatopathology specialty (covariate coded 1,0) and actual performance (continuous independent variable) in separate general linear regression models (analysis of covariance), for both pre- and post-educational interventions. Least square mean estimation was used to compare the mean difference between perceived and actual performance at the mean value of the covariate. Inference of statistical significance was based on an alpha level of 0.05 or equivalently, when the 95% CI crosses zero. Analyses were performed using a commercially available SAS v9.4 (SAS Institute, Cary NC).

To assess the effect participation in both educational interventions may have had on study results, we conducted a sensitivity analysis by removing data on the four pathologists who participated in both educational interventions from the intervention data for the skin educational intervention, because this intervention occurred second, about 2 years after the breast intervention. We found no differences in study findings when they were in or out of the skin intervention analyses, so these pathologists were included in the analyses of both interventions.

Results

The characteristics of the pathologists who interpret breast biopsy specimens were similar with regard to age, gender, affiliation with an academic medical center, and years interpreting breast and/or melanocytic lesions to those who interpreted skin biopsy

specimens (Tables 1a and 1b). Pathologists who interpret skin biopsies were more likely to be fellowship trained or board certified than those who interpret breast biopsies (3% for breast and 40% for dermatopathologists, $p<.001$; Table 1a). Pathologists interpreting skin tissue reported signing out more cases per month than those interpreting breast biopsies (28% having >150 cases per month versus 2%, $p<.001$; Table 1a), and 44% of skin pathologists versus 18% of breast pathologists perceived they were considered experts by their peers ($p<.001$). Additionally, 87% of board certified dermatopathologists were self-reported to be considered an expert by their peers. Ninety-five percent of skin pathologists reported finding pathological interpretation of tissue challenging versus 52% of breast pathologists ($p<.001$). Similarly, 70% of skin pathologists reported that interpreting melanocytic cutaneous lesions makes them more nervous than other types of pathology (with skin pathologists with additional certification/fellowship training in dermatopathology even more so at 76%), whereas 48% of breast pathologists reported that interpreting breast tissue makes them more nervous than other types of tissue ($p<.001$; Table 1b). An equivalent amount of median time spent on the CME website was reported by skin pathologists and breast pathologists (2 hours vs. 2 hours, $p=0.29$) and median time spent by pathologists in both groups to interpret the initial slide set was also similar (8 hours vs. 8 hours, $p=0.45$).

The mean pre-intervention perceived agreement for all pathologists interpreting skin tissue was 72.9%, while the actual agreement was 52.4% (Figure 2a-1). Among dermatopathologists, the mean perceived agreement was 76.9% prior to the educational intervention, while the actual agreement was 59.5%, an overestimated mean difference of 17.4% (95% CI 11.8% to 23.0%) (Figure 2a-2). Among pathologists without specialty training who interpreted skin tissue at pre-intervention (Figure 2a-3), the mean perceived agreement was 70.3%, while the actual agreement was 47.7%, an overestimated mean difference of 22.6% (95% CI 18.2% to 27.0%). The mean difference in perceived versus actual agreement between the two dermatopathology subgroups did not significantly differ overall ($p=0.13$) or separately within each of five diagnostic classes (all p -values well over 0.05) either prior to or after the educational intervention, with one exception. For cases given a Class III assignment by expert consensus, when compared to the pathologists' actual performance, other skin pathologists overestimated their perceived performance by 11.9% (95% CI 3.0% to 20.8%) more on average than pathologists who were dermatopathology certified and/or fellowship trained prior to the educational intervention and by 8.9% (95% CI 2.6% to 15.3%) post-educational intervention (data not shown). Overall mean perceived agreement for pathologists interpreting breast tissue (the extent they thought they would agree with the reference diagnosis) was 81.4% (all four categories combined), while the actual agreement was 75.9% (Figure 2b).

Figure 3a-1 illustrates the alignment among all skin pathologists for their five diagnostic categories pre- and post-intervention. As shown, alignment among these pathologists was much lower for nevus, mild atypia (Class I), moderate atypia (Class II), and melanoma *in situ* (Class III) at about 60%. Figure 3a-2 shows this alignment among dermatopathologists (those with fellowship training/board certification) pre- and post-intervention. Those with additional training/credentialing were only slightly more aligned than this group of pathologists as a whole where nevus, mild atypia (Class I) and moderate atypia (Class II)

aligned at about 66-68%, and melanoma *in situ* (Class III) was at about 60% and alignment for pT1a invasive melanoma and pT1b were about 45%. Figure 3a-3 illustrates this alignment among pathologists without fellowship training pre- and post-intervention. Alignment was highest at about 66% for melanoma *in situ* and was lowest for invasive melanoma (both categories) at about 45-47%.

Figure 3b shows how well the intervention aligned breast pathologists' perceptions with agreement with the reference consensus diagnosis for each diagnostic category pre- and post-intervention assessment, illustrating the majority of participants both processed and recalled what they learned about the extent of agreement with the reference diagnosis. In the first three panels where the reference diagnoses were benign without atypia, atypia, and DCIS, the alignment reached 82-88%, whereas the alignment for invasive reached 93%.

Discussion

This study is the first to our knowledge to explore how comparisons of pathologists interpreting lesions from two different organ systems, breast and skin, might differ and how these pathologists perceived their diagnoses would agree with a reference diagnosis determined by a consensus panel of experienced pathologists. Our intent was to examine what might be learned from comparing and contrasting these complex interpretive experiences rather than attempting to conform interpretive practices from each into one standardized approach, which should not be done.

We learned that the magnitude of difference among pathologists interpreting skin lesions compared to those interpreting breast tissue perceptions of diagnostic agreement was approximately four times greater, indicating they perceived they would agree with the reference diagnosis much more often than they actually did. In addition, after undertaking the intervention and reviewing cases where they both agreed with and did not agree with the reference diagnosis and received case-specific microscopic diagnostic comments about each case, skin pathologists were less likely to process or recall alignment with the reference diagnosis compared to breast pathologists. Surprisingly, while 95% of those interpreting breast tissue were especially likely to align with the reference diagnosis at the end of the intervention for invasive cancer, both invasive categories for those interpreting skin tissue only aligned at about 45%, which was lower than benign or *in situ* cases.

There are a few reasons why this might have occurred. One is that very few of the participating pathologists participating in the breast study had fellowship training, as breast fellowships are not as common as those in dermatopathology. It could be that this lack of additional training resulted in enhanced uptake of the intervention because these pathologists were less confident in their interpretations and perhaps were more open to learning about unanticipated nuances of breast pathology interpretation. Our study also shows that pathologists who interpret skin biopsies, on average, see a significantly larger number of cases than pathologists who interpret breast biopsies for any given time period, which may have also affected the uptake of the intervention among those interpreting skin tissue, who may be more confident of their interpretations. Our findings may also be the result of

differences in the background subjectivity of melanoma and breast pathology. For example, in skin interpretation, criteria are broader which could lead to more interpretive subjectivity.

Despite additional training and being exposed to more cases regularly, pathologists interpreting skin tissue, even those with fellowship training, were considerably less accurate in estimating their level of agreement with the reference diagnosis after the educational intervention than their colleagues interpreting breast pathology. Another possible explanation is the presence of multiple philosophical differences among dermatopathologists, likely based on where they received fellowship training. For example, not all dermatopathologists believe in the grading of dysplasia (17) while others aggressively dismiss the idea of dysplasia altogether (18). These differences have resulted in a highly complex system of classification for melanocytic lesions with the development of local variations in terminology and criteria that make agreement across different centers more difficult to achieve, perhaps, than in breast cancer where there is more standardization. Regardless of philosophical stance, dermatopathologists performance was more variable in our model. One key factor missing from all approaches to the melanocytic lesion is the difficulty of correlation with the true biologic nature of the lesion. A melanocytic lesion that appears atypical to some pathologists may not appear atypical to others. This is also true for borderline breast lesions and even small foci of DCIS.

Borderline cases generate disagreement and differences of opinion among pathologists. Our own research (13) has shown that accuracy using a consensus diagnosis of experienced pathologists as reference varied by class with agreement for nevus or mild atypia at 92% (95% confidence interval 90% to 94%); moderate atypia at 25% (22% to 28%); severe atypia or melanoma *in situ* at 40% (37% to 44%); early invasive melanoma at 43% (39% to 46%); and invasive melanoma at 72% (69% to 75%). For biologically indolent lesions, these classification disagreements do not generally result in clinically catastrophic events that lead pathologists to systematically change diagnostic classification systems. However, borderline indolent cases are at the root of the current discussion and evaluation of over diagnosis and over treatment. The use of second opinions have implications for these findings, and we are planning a separate modeling sub-study to specifically examine the use of second opinions in clinical practice.

Nonetheless, significant experience and acumen in this field may provide insight into the appropriate recognition and management of many ambiguous/uncertain melanocytic lesions. Currently there are no genetic tests, immunohistochemical markers or any other test that can help differentiate who is closer to the biological verities, which are largely opaque to observers except in the event of metastatic disease. Standardization of the approach, such as the use of MPATH-Dx© (19), would be an important first step in eliminating the myriad of terminologies in existence and embracing the fact that there is much to be learned about the sub-classification of melanocytic skin tumors, especially those in the “intermediate” categories between benign lesions and melanomas, and also the low risk melanomas, which may have capacity for local persistence, recurrence and progression, but lack capacity for metastasis.

In any case, this study has important implications for the new Maintenance of Certification (MOC) standards that were implemented in January of 2015 (20), especially for Part II which involves self-monitoring and life-long learning. The rationale for Part II standards include the need for MOC to contribute to better patient care by requiring ongoing participation in high quality, unbiased learning and self-assessment activities relevant to specialty and areas of practice (20). We endeavored to foster the development of un-biased consensus diagnoses in both these educational interventions, and they specifically included rigorous self-assessment directly related to clinical pathology. In addition, Part II MOCs are intended to be relevant, easy-to-use, cost-effective, and meaningful for practicing physicians and, given our high rates of completion, we believe that we achieved these additional outcomes. The new MOC system is not without controversy (21) with many physicians raising concerns about expense, burden, and whether it is a clinically relevant process. What is needed in the new MOC system is evidence that patient care is better as a result, which is currently lacking. Our intervention was not directly associated with patient outcomes. This type of work must be done to provide this needed evidence.

Another relevant area for implications involves malpractice. We published two studies examining pathologists' perspectives on how concerns about medical malpractice may affect their interpretive practices. In both papers (22, 23) the majority of pathologists reported undertaking protective behaviors due to concerns about medical **malpractice**, including ordering additional stains, recommending additional surgical sampling, obtaining second reviews, or choosing the more severe diagnosis for borderline cases. More research should be done in this area.

The strengths of this study include the representation by about a third of eligible breast **pathologists** and just over half of eligible dermatopathologists from several states in the U.S. In addition, comparisons were made using two sets of cases with very carefully developed reference diagnoses for both breast and skin tissues. Weaknesses include that interpretations involved only one slide per case, which is not representative of actual clinical practice. Additionally, the skin pathologists reviewed melanocytic skin cases rather than general skin cases. As a result, the skin cases included in the test sets were more difficult lesions to interpret. Another weakness is that pathologists in this study likely interpret both breast and skin tissue, and we do not know the extent that this is the case among participants. The educational interventions also included subtle wording differences, which could have affected perceptions of agreement. For example, in the Breast CME, the reference panel was referred to as "the expert panel" and in the skin CME, the reference panel was referred to as "the consensus panel." We do not expect such differences would have affected the results to a great extent. Lastly, there were scale differences between the two studies in that the skin intervention used five diagnostic categories, while the breast intervention used four. These differences were not addressed statistically, although participants in both interventions provided an estimate between 0%-100%. However, the scale differences may have influenced the choices they made. Also, the summary rates of agreement are compared ignoring the denominator within and between M-Path (36 vs. 48 cases) and B-Path studies (60 cases).

In conclusion, pathologists interpreting breast tissue appeared to experience higher comprehension of their performance of the intervention based on their alignment with the consensus diagnosis after the intervention compared to pathologists interpreting skin biopsies.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers* R01 CA140560, R01 CA151306, R01 CA172343 and K05 (CA104699). The content is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health.

References

1. Braun RP, Gutkowitz-Krusin D, Rabinovitz H, Cagnetta A, Hofmann-Wellenhof R, Ahlgrimm-Siess V, Polsky D, Oliviero M, Kolm I, Googe P, King R, Prieto VG, French L, Marghoob A, Mihm M. Agreement of dermatopathologists in the evaluation of clinically difficult melanocytic lesions: how golden is the 'gold standard'? *Dermatology*. 2012; 224(1):51–58. [PubMed: 2243231]
2. Elmore JG, Nelson HD, Pepe MS, Longton G, Tosteson ANA, Geller BM, Onega T, Carney PA, Jackson SL, Allison KH, Weaver DL. Variability in Pathologists' Interpretations of Individual Breast Biopsy Slides: A Population Perspective. *Annals of Int Med*. 2016; 164(10):649–655.
3. Elmore JG, Longton G, Carney PA, Geller BM, Onega T, Tosteson ANA, Nelson H, Pepe M, Allison KH, Schnitt S, O'Malley F, Weaver DL. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA*. 2015; (11):1122–1132. [PubMed: 25781441]
4. Elmore JG, Tosteson ANA, Pepe M, Longton G, Nelson HD, Geller BM, Carney PA, Onega T, Allison KH, Jackson S, O'Malley F, Weaver DL. Evaluation of 12 strategies for obtaining second opinions to improve interpretation of breast histopathology: simulation study. *BMJ*. 2016; 353:i3069. [PubMed: 27334105]
5. Carney PA, Reisch LM, Piepkorn MW, Barnhill RL, Elder DE, Knezevich S, Longton G, Elmore JG. Achieving Consensus for the Histological Diagnosis of Melanocytic Lesions: Use of the Modified Delphi Method. *J Cutaneous Path*. 2016; 43(10):830–837.
6. Knezevich S, Barnhill RL, Elder DE, Piepkorn MW, Reisch LM, Pocobelli G, Carney PA, Elmore JG. Variability in mitotic figures in serial sections of thin melanomas. *JAAD*. 2014; 71(6):1204–11. [PubMed: 25239732]
7. Puliti D, Duffy S, Miccinesi G, de Koning H, Lynge E, Zappa M, Paci E. Methodology and estimate of overdiagnosis in breast cancer service screening: a review of the European studies. *J Med Screen*. 2012; 19(S1):42–54.
8. Weyers W. The 'epidemic' of melanoma between under- and over-diagnosis. *J Cutaneous Path*. 2012; 39(1) <http://dx.doi.org/10.1111/j.1600-0560.2011.01831.x>.
9. Elmore JG, Tosteson ANA, Pepe M, Longton G, Nelson HD, Geller BM, Carney PA, Onega T, Allison KH, Jackson S, O'Malley F, Weaver DL. Evaluation of 12 strategies for obtaining second opinions to improve interpretation of breast histopathology: simulation study. *BMJ*. 2016; 353:i3069. [PubMed: 27334105]
10. Nakhleh RE, Bekeris LG, Souers RJ, Meier FA, Tworek JA. Surgical pathology case reviews before sign-out: a College of American Pathologists Q-Probes study of 45 laboratories. *Arch Pathol Lab Med*. 2010; 134:740–743. [PubMed: 20441505]
11. Geller BM, Nelson HD, Carney PA, Weaver DL, Onega T, Allison KH, Frederick PD, Tosteson AT, Elmore JG. Second Opinion in Breast Pathology: Policy, Practice and Perception. *J Clin Pathol*. 2014; 0:1–6.
12. Wick MR, Ocknew DM, Mills SE, Ritter JH, Swanson PE. Homologous Carcinomas of the Breasts, Skin, and Salivary Glands: A Histologic and Immunohistochemical Comparison of Ductal Mammary Carcinoma, Ductal Sweat Gland Carcinoma, and Salivary Duct Carcinoma. *Am J Clin Pathol*. 1998; 109(1):75–84. [PubMed: 9426521]

13. Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, Carney PA, Titus LJ, Nelson HD, Onega T, Tosteson ANA, Weinstock MA, Knezevich SR, Piepkorn MW. The Reproducibility and Accuracy of Pathologists' Diagnosis of Invasive Melanoma and Melanocytic Proliferations. *BMJ*. 2017; 357:j2813. [PubMed: 28659278]
14. Oster N, Carney PA, Allison KH, Weaver D, Reisch L, Longton G, Onega T, Pepe M, Geller BM, Nelson H, Ross T, Elmore JG. Development of a diagnostic test set to assess agreement in breast pathology: Practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *BMC Women's Health*. 2013; 13:3. [PubMed: 23379630]
15. Carney PA, et al. The New Hampshire Mammography Network: the development and design of a population-based registry. *AJR*. 1996; 167(2):367–72. [PubMed: 8686606]
16. Breast Cancer Surveillance Consortium. Available at: <http://breastscreening.cancer.gov/>, (Accessed 9/29/14).
17. Lozeau DF, Farber MJ, Lee JB. A nongrading histologic approach to Clark (dysplastic) nevi: A potential to decrease the excision rate. *J Am Acad Dermatol*. 2016; 74(1):68–74. [PubMed: 26549250]
18. Ackerman AB. What naevus is dysplastic, a syndrome and the commonest precursor of malignant melanoma? A riddle and an answer. *Histopathology*. 1988; 13(3):241–56. [PubMed: 3056824]
19. Piepkorn MW, Barnhill RL, Elder DE, Knezevich SR, Carney PA, Reisch LM, Elmore JG. The MPATH Reporting Schema for Melanocytic Proliferations and Malignant Melanoma. *JAAD*. 2014; 70(1):131–141. [PubMed: 24176521]
20. American Board of Medical Specialties. Standards for the ABMS Program for Maintenance of Certification (MOC). <http://www.abms.org/media/1109/standards-for-the-abms-program-for-moc-final.pdf> (Accessed 6/27/17).
21. Lowes, R. ABIM Suspends Controversial MOC Requirements Through 2018. *Medscape*. Dec 16, 2015. <http://www.medscape.com/viewarticle/856076> (accessed 6/27/17).
22. Reisch LM, Carney PA, Oster NV, Weaver DL, Nelson HD, Frederick PD, Elmore JG. Medical Malpractice Concerns and Defensive Medicine: A Nation-wide Survey of Breast Pathologists. *Am J Clin Pathol*. 2015 Dec; 144(6):916–22. [PubMed: 26572999]
23. Carney PA, Fredricks PD, Reisch LM, Knezevich S, Piepkorn MW, Barnhill RL, Elder DE, Beller BM, Titus L, Weinstock MA, Nelson HD, Elmore JG. How Concerns about Medical Malpractice Affect Pathologists' Perceptions of their Diagnostic Practices when Interpreting Melanocytic Lesions. *JAAD*. 2016; 74(2):317–324. [PubMed: 26559597]

Figure 1. Inclusion of Pathologists Who Interpret Skin and Breast Biopsies

^aincluding 3 with fellowship training

^b7 did not participate in CME & 4 provided incomplete data when estimating performance

^c7 did not participate in CME & 11 provided incomplete data when estimating performance

^d21 did not complete CME and 2 did not provide complete performance estimates

Step I: Participants Interpret Test Sets on either Skin and Breast Cases

Step II: CME Program Components (*Web-hosted & Tailored to Individual Participants upon Log-in*)

Pre-intervention

1. Participants indicate, before beginning the program, how they believe their interpretations will compare in terms of agreement with the reference diagnoses overall and for each diagnostic category.

Intervention

2. Participants review:
 - a. How each of their case interpretations compare to reference the diagnoses.
 - b. Digital images of slides with annotated microscopic diagnostic comments for cases they interpreted both correctly and incorrectly for as long as desired.

Post-Intervention

3. Participants indicate, after completing the program, how their interpretations aligned with the reference diagnosis to assess uptake of the intervention.

Figure 2. Study Activities Related to the CME Program

(a-1) Comparison of Perceived versus Actual Agreement Among All Skin Pathologists at the Pre-Intervention Assessment (n=158)

*From matched-pairs t-test for the difference between the pair of means expressed as mean difference (95% confidence interval)

(a-2) Comparison of Perceived versus Actual Agreement Among Fellowship-Trained and/or Board Certified Dermatopathologists at the Pre-Intervention Assessment (n=63)

* From matched-pairs t-test for the difference between the pair of means expressed as mean difference (95% confidence interval)

a-3. Comparison of Perceived versus Actual Agreement Among Other Skin Pathologists with No Special Training in Dermatopathology at the Pre-Intervention Assessment (n=95)

* From matched-pairs t-test for the difference between the pair of means expressed as mean difference (95% confidence interval)

a.-2.

Figure 3.

a-1. Differences in Perceived versus Actual Agreement Among All Skin Pathologists as Part of the Educational Intervention According to Individual Diagnostic Categories (Pre-and Post-intervention) (n=158)

* The proportion of participants whose difference in perceived versus actual agreement fell within plus-minus 5% of zero difference

(a.-2) Differences in Perceived versus Actual Agreement Among Fellowship-Trained and/or Board Certified Dermatopathologists as Part of the Educational Intervention According to Individual Diagnostic Categories (Pre- and Post-intervention) (n=63)

*The proportion of participants whose difference in perceived versus actual agreement fell within plus-minus 5% of zero difference

(a.-3) Differences in Perceived versus Actual Agreement Among Other Skin Pathologists with No Special Training in Dermatopathology as Part of the Educational Intervention According to Individual Diagnostic Categories (Pre- and Post-intervention) (n=95)

* The proportion of participants whose difference in perceived versus actual agreement fell within plus-minus 5% of zero difference

(b) Differences in Perceived versus Actual Agreement Among Breast Pathologists as Part of the Educational Intervention According to Individual Diagnostic Categories (Pre- and Post-intervention) (n=92)

* The proportion of participants whose difference in perceived versus actual agreement fell within plus-minus 5% of zero difference

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1a

Breast, Skin, and Dermatopathologists Demographic, Training and Practice Characteristics

	Pathology Specialties				P values ^a		
	M-Path Study		B-Path Study		1 vs 4	2 vs 3	3 vs 4
Participant Characteristics	All Skin Pathologists [1]	Dermatopathology Certified and/or Fellowship Trained [2]	No Specialty Training in Dermatopathology [3]	Breast Pathologists [4]			
Total	158 (39)	63 (15)	95 (23)	92 (23)			
Demographics							
Age (yrs.)					0.58	<.001	0.030
< 40	23 (15)	18 (29)	5 (5)	13 (14)			
40-49	52 (33)	24 (38)	28 (29)	32 (35)			
50-59	52 (33)	16 (25)	36 (38)	35 (38)			
60	31 (20)	5 (8)	26 (27)	12 (13)			
Gender					0.93	0.27	0.68
Female	61 (39)	21 (33)	40 (42)	36 (39)			
Male	97 (61)	42 (67)	55 (58)	56 (61)			
Training and Experience							
Affiliation with Academic Medical Center					0.88	<.001	0.043
No	114 (72)	35 (56)	79 (83)	67 (73)			
Yes, adjunct/ affiliated	29 (18)	15 (24)	14 (15)	15 (16)			
Yes, primary appointment	15 (9)	13 (21)	2 (2)	10 (11)			
Training ^b					<.001	<.001	0.12
Trained pathology	63 (40)	63 (100)	0 (0)	3 (3)			
In dermatopathology	95 (60)	0 (0)	95 (100)	89 (97)			
Years Interpreting Specimens					0.95	<.001	0.25
< 10	63 (40)	36 (57)	27 (28)	35 (38)			
10-19	48 (30)	20 (32)	28 (29)	28 (30)			
20	47 (30)	7 (11)	40 (42)	29 (32)			
Percent of Caseload ^c					0.14	<.001	0.049
< 10	67 (42)	4 (6)	63 (66)	48 (52)			

Participant Characteristics	Pathology Specialties				P values ^a		
	All Skin Pathologists [1]	Dermatopathology Certified and/or Fellowship Trained [2]	No Specialty Training in Dermatopathology [3]	B-Path Study Breast Pathologists [4]	1 vs 4	2 vs 3	3 vs 4
10	91 (58)	59 (94)	32 (34)	44 (48)			
<i>Average Number of Total Cases per Month ^d</i>							
50	65 (41)	3 (5)	62 (65)	58 (63)			
50-140	49 (31)	26 (41)	23 (24)	32 (35)			
150	44 (28)	34 (54)	10 (11)	2 (2)			
<i>Considered an expert by colleagues</i>							
No	89 (56)	8 (13)	81 (85)	75 (82)	<.001	<.001	0.49
Yes	69 (44)	55 (87)	14 (15)	17 (18)			

Note: percentages may not total 100 due to rounding

^a As a correction for multiple comparisons, alpha level = 0.017.

^b Training: Trained in breast pathology **or** breast and surgical pathology (vs not (includes surgical pathology)) or for dermatopathologists, either fellowship trained or board certified in dermatopathology (vs fellowship or board certified in other topic).

^c For breast, percent of caseload interpreting breast specimens; for dermpath, percent of usual caseload of MSL

^d For breast, original categorical variable in weeks converted to months; for dermpath, total benign MSL and cases of melanoma (including MIS and invasive melanoma) interpreted per month then categorized

Table 1b

Breast, Skin, and Dermatopathologists Demographic, Training and Practice Characteristics

Participant Characteristics	Pathology Specialties				P values ^a		
	All Skin Pathologists [1]	Dermatopathology Certified and/or Fellowship Trained [2]	No Specialty Training in Dermatopathology [3]	B-Path Study	1 vs 4	2 vs 3	3 vs 4
Total	158 (39)	63 (15)	95 (23)	Breast Pathologists [4]			
				92 (23)			
Perceptions about Interpretation ^e							
In general, how challenging do you find pathological interpretation of biopsy specimens?							
Challenging	150 (95)	59 (94)	91 (96)	48 (52)	<.001	0.71	<.001
Not Challenging	8 (5)	4 (6)	4 (4)	44 (48)			
Interpreting breast or melanocytic skin lesions makes me more nervous than other types of pathology							
Nervous	111 (70)	48 (76)	63 (66)	44 (48)	<.001	0.18	0.011
Not Nervous	47 (30)	15 (24)	32 (34)	48 (52)			
In general, how confident are you in your assessments of of biopsy specimens?							
Low confidence	21 (13)	5 (8)	16 (17)	8 (9)	0.27	0.11	0.096
Confident	137 (87)	58 (92)	79 (83)	84 (91)			
Interpreting biopsy specimens							
Not enjoyable	42 (27)	10 (16)	32 (34)	21 (23)	0.51	0.013	0.10
Enjoyable	116 (73)	53 (84)	63 (66)	71 (77)			
How similar were the types of cases included in the CME glass slides you interpreted compared to the entire spectrum of lesions you see in your practice? ^f							
Never (rarely) see cases like these	4 (3)	0 (0)	4 (4)	0 (0)	<.001	0.003	<.001
Sometimes see cases like these	55 (35)	13 (21)	42 (45)	21 (24)			
Often see cases like these	61 (39)	28 (44)	33 (35)	46 (52)			
Almost always see cases like these	18 (11)	11 (17)	7 (7)	0 (0)			
Always see cases like these	19 (12)	11 (17)	8 (9)	22 (25)			
Missing	1 (1)	0 (0)	1 (1)	3 (3)			
Which of the following types of CME do you most prefer? ^g							
Instructor-led programs	73 (46)	29 (46)	44 (46)	60 (65)	<.001	0.21	0.021
Self-directed programs	75 (47)	34 (54)	41 (43)	24 (26)			

Participant Characteristics	Pathology Specialties				P values ^a		
	M-Path Study		B-Path Study		1 vs 4	2 vs 3	3 vs 4
	All Skin Pathologists [1]	Dermatopathology Certified and/or Fellowship Trained [2]	No Specialty Training in Dermatopathology [3]	Breast Pathologists [4]			
Other	4 (3)	0 (0)	4 (4)	7 (8)			
Missing	6 (4)	0 (0)	6 (6)	1 (1)			
Estimate the amount of time you spent completing surveys and reviews of slides (hours)							
Reviewing CME website							
Hours, median (Q1, Q3)	2 (1, 3)	2 (1, 2)	2 (1, 3)	2 (1, 5)	0.12	0.34	0.29
Total hours ^h							
Hours, median (Q1, Q3)	20 (11, 20)	19 (10, 20)	20 (12, 20)	17 (11, 20)	0.77	0.37	0.45

Note: percentages may not total 100 due to rounding

^a As a correction for multiple comparisons, **alpha** level = 0.017.

^e original 6-pt Likert scale variables dichotomized.

^f 'never or rarely see cases like these' and 'never see cases like these' was provided to M-Path and B-Path participants, respectively, and are treated as equal responses; 'almost always see cases like these' was not a response provided in B-Path study but was provided in M-Path study. Missing value not included in inferential statistics.

^g Self-directed programs that are online and self-directed programs that are not online were possible responses in M-Path study and were combined to match the response provided in B-Path (self-directed responses). Missing value not included in inferential statistics.

^h Total hours includes time spent filling out baseline survey in addition to the time spent reviewing slides during the two phases of the study and the CME website.