Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study

Joann G Elmore,[1] Raymond L Barnhill,[2] David E Elder,[3] Gary M Longton,[4] Margaret S Pepe,[4] Lisa M Reisch,[1] Patricia A Carney,[5] Linda J Titus,[6] Heidi D Nelson,[7,8] Tracy Onega,[9,10] Anna N A Tosteson,[11] Martin A Weinstock,[12,13] Stevan R Knezevich,[14] Michael W Piepkorn[15,16]

For numbered affiliations see end of article.

Correspondence to: J G Elmore jelmore@uw.edu

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2017;357:j2813 http://dx.doi.org/10.1136/bmj.j2813

Accepted: 25 May 2017

ABSTRACT

OBJECTIVE

To quantify the accuracy and reproducibility of pathologists' diagnoses of melanocytic skin lesions.

DESIGN

Observer accuracy and reproducibility study.

SETTING

10 US states.

PARTICIPANTS

Skin biopsy cases (n=240), grouped into sets of 36 or 48. Pathologists from 10 US states were randomized to independently interpret the same set on two occasions (phases 1 and 2), at least eight months apart.

MAIN OUTCOME MEASURES

Pathologists' interpretations were condensed into five classes: I (eg, nevus or mild atypia); II (eg, moderate atypia); III (eg, severe atypia or melanoma in situ); IV (eg, pathologic stage T1a (pT1a) early invasive melanoma); and V (eg, ≥pT1b invasive melanoma). Reproducibility was assessed by intraobserver and interobserver concordance rates, and accuracy by concordance with three reference diagnoses.

RESULTS

In phase 1, 187 pathologists completed 8976 independent case interpretations resulting in an average of 10 (SD 4) different diagnostic terms applied to each case. Among pathologists interpreting the same cases in both phases, when pathologists diagnosed a case as class I or class V during phase 1, they gave the same diagnosis in phase 2 for the majority of cases (class I 76.7%; class V 82.6%). However, the intraobserver reproducibility was lower for cases interpreted as class II (35.2%), class III (59.5%), and class IV (63.2%). Average interobserver concordance rates were lower, but with similar trends. Accuracy using a consensus diagnosis of experienced pathologists as reference varied by class: I, 92% (95% confidence interval 90% to 94%); II, 25% (22% to 28%); III, 40% (37% to 44%); IV, 43% (39% to 46%); and V, 72% (69% to 75%). It is estimated that at a population level, 82.8% (81.0% to 84.5%) of melanocytic skin biopsy diagnoses would have their diagnosis verified if reviewed by a consensus reference panel of experienced pathologists, with 8.0% (6.2% to 9.9%) of cases overinterpreted by the initial pathologist and 9.2% (8.8% to 9.6%) underinterpreted.

CONCLUSION

Diagnoses spanning moderately dysplastic nevi to early stage invasive melanoma were neither reproducible nor accurate in this large study of pathologists in the USA. Efforts to improve clinical practice should include using a standardized classification system, acknowledging uncertainty in pathology reports, and developing tools such as molecular markers to support pathologists' visual assessments.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Millions of skin biopsy samples are obtained each year

A pathologist's visual interpretation is the cornerstone for diagnosing melanocytic lesions, including melanoma, yet previous studies have suggested variability among pathologists in their diagnoses

WHAT THIS STUDY ADDS

Diagnoses within the disease spectrum from moderately dysplastic nevi to early stage invasive melanoma are neither reproducible nor accurate

These limitations in histological diagnosis emphasize the need for supplemental reporting paradigms to convey observer derived opinions about diagnostic uncertainty, perceived risk for disease progression, and suggested management

Use of a standardized classification format employing unambiguous language and acknowledging uncertainty in pathology reports might reduce the potential for miscommunication and management errors

Introduction

Diagnostic errors contribute to approximately 10% of patient deaths and are the top cause of medical malpractice payouts.[1] A recent report by the National Academies of Sciences, Engineering, and Medicine deemed improvements in the diagnostic process "a moral, professional, and public health imperative."[1]

The goal of a medical diagnosis is to identify and assign natural phenomena to the correct diagnostic classification with both accuracy and precision. However, inadequate development and inconsistent use of disease labels and classification schemes by clinicians can lead to patient harm.[2 3] While physicians may observe similar features on a biopsy sample slide or radiograph or on a patient's physical examination, their diagnosis reflects individual perspectives in the processing, assigning of importance, and categorizing of medical information. As diagnostic criteria increase in their subjectivity, diagnoses between physicians become increasingly discordant.

With the escalating incidence of melanoma now exceeding the rates of increase of all other major cancers,[4] the diagnosis of cutaneous melanocytic lesions exemplifies the challenges physicians face when interpreting and classifying medical data. The diagnosis of cutaneous melanocytic lesions relies on a pathologist's visual assessment of biopsy material on microscopic slides. The reliability and predictive values of the

diagnostic criteria used for these lesions have never been established with rigorous standards. Previous studies have suggested high levels of diagnostic discordance among pathologists in the interpretation of melanocytic lesions,[5-8] alluding to the potential for both overdiagnosis and underdiagnosis,[4 9] yet these older studies were limited in number of specimens and pathologists. Thus there is a critical need to evaluate the current quality of diagnostic practices in this specialty and consider the impact at the population level.

We evaluated the reproducibility and accuracy of melanocytic lesion diagnoses provided by a broad spectrum of practicing pathologists in the USA. Reproducibility was assessed by intraobserver and interobserver concordance rates. Accuracy was evaluated using three different reference diagnoses. We then estimated how diagnostic variability affected accuracy from the perspective of an adult patient in the USA. These evaluations are overdue because of the therapeutic ramifications, emotional and physical burdens of diagnosis, and utilization of healthcare resources.

Methods

Study procedures included recruitment of pathologists in several US states who interpret melanocytic lesions, a baseline survey, independent interpretations of melanocytic lesions at two time points, a personalized feedback module, and a post-interpretation survey. Detailed information about development of the cases, participant recruitment, and diagnostic classification is provided elsewhere.[10-14]

Histology form and mapping scheme

We gathered assessment and recommendations on each case using a standardized online histology form and then classified the diverse terms using the Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx) histology form.[13] This tool organizes pathologists' diverse descriptive terms of melanocytic skin lesions into five diagnostic classes, with suggested treatment recommendations (table 1). Example diagnostic terms for each class (and suggested treatment recommendations, all provided under the assumption that specimen margins are positive) include: class I, nevus or mild atypia (no further treatment margin required); class II, moderate atypia (consider narrow but complete repeat excision margin <5 mm); class III, severe atypia or melanoma in situ (repeat excision with ≥5 mm but <1 cm margins); class IV, pT1a invasive melanoma (wide excision ≥1 cm margin); and class V, ≥pT1b invasive melanoma (wide excision ≥1 cm with possible additional treatment, such as sentinel lymph node biopsy and adjuvant therapy).

Biopsy case development

Cutaneous melanocytic lesions from shave, punch, and excisional specimens were included.[13 14 16] We selected cases using stratification based on patient age (20-49 years, 50-64 years, ≥65 years) and medical chart documentation of the original diagnosis. Three experienced dermatopathologists with recognized expertise in cutaneous melanocytic lesions (RLB, DEE, MWP) prepared and independently reviewed new haematoxylin and eosin stain glass slides for each case, followed by consensus review using a modified Delphi approach.[14 17] The final 240 cases had intentionally higher proportions in classes II-V than are typically encountered in practice: 10.4% (n=25) in class I, 15.0% (n=36) in class II, 25.0% (n=60) in class III, 24.2% (n=58) in class IV, and 25.4% (n=61) in class V. We assembled the 240 cases into five sets of 48 cases. All participants interpreted 48 cases in phase 1. In phase 2, those who agreed to alternatively participate in a substudy of digital whole slide imaging interpreted a subset of 36 of their original 48 cases; otherwise, participants who declined the substudy interpreted the same 48 cases as in phase 1 (see figure 1 in web appendix).

Pathologist identification, recruitment, and baseline characteristics

We identified potential participants using publicly available information from 10 US states (CA, CT, HI, IA, KY, LA, NJ, NM, UT, and WA). Pathologists were invited by email, postal mail, and telephone. Eligible participants had completed their pathology residencies and/or fellowships, interpreted cutaneous melanocytic lesions within the previous year, and expected to interpret cutaneous melanocytic lesions for the next two years. Participants completed a baseline survey assessing their personal and clinical practice characteristics.[10 12] To compare characteristics among participants and non-participants, we obtained additional information from Direct Medical Data.[18]

Table 1 | Summary of MPATH-Dx reporting schema for classification of melanocytic skin lesions into five diagnostic categories

MPATH-Dx class	Perceived risk for progression	Suggested intervention*	Examples
0	Incomplete study due to sampling or technical limitations	Repeat biopsy or short term follow-up	NA
I	Very low risk	No further treatment	Common melanocytic nevus; blue nevus; mildly dysplastic nevus
II	Low risk	Narrow but complete excision (<5 mm)	Moderately dysplastic nevus; -Spitz nevus
III	Higher risk. Greater need for intervention	Complete excision with at least 5 mm but <1 cm margins	Severely dysplastic nevus; melanoma in situ; atypical Spitz tumor
IV	Substantial risk for local or regional progression	Wide local excision with ≥1 cm margins	Thin, invasive melanomas (eg, pT1a†)
V	Greatest risk for regional and/or distant metastases	Wide local excision with ≥1 cm margins. Consideration of staging sentinel lymph node biopsy, adjuvant therapy	Thicker, invasive melanomas (eg, pT1b, stage 2 or greater†)

*Assuming representative sampling of lesion.
†According to American Joint Committee on Cancer seventh edition cancer staging manual.[15]

Interpretations by participants

We stratified randomization to specific slide sets by pathologists' clinical expertise, which was dichotomized according to possession of one or more of the following self reported characteristics on the baseline survey: fellowship trained or board certified in dermatopathology; considered by colleagues to be an expert in melanocytic lesions; or 10% or more of usual caseload included cutaneous melanocytic lesions.

Slides were presented in a random order to each participant. Participants were provided with the patient's age, biopsy type, and anatomic location of biopsy site. Standardized diagnostic definitions or educational materials were not provided. We asked the pathologists to assume that the single glass slide was representative of the entire lesion, and that the margin was involved (irrespective of whether it involved the biopsy margin or not). Participants were asked to complete their interpretations within one week.

Pathologists provided their diagnoses using the online histology data collection form, which contained 56 different terms[13] (see table 1 in web appendix). For each case we summed and averaged the total number of different diagnosis terms used to describe that case in phase 1. Eight or more months after completing initial interpretations in phase 1, pathologists assessed the same cases a second time in phase 2. For the second phase, we presented the cases in a different randomly assigned order for each participant; pathologists were not informed that these were the same cases. As the same glass slide was used for each case, only one participant could have the test sets at a time. Thus, more than three years were required for data collection from all participants on the baseline survey, the phase 1 and 2 data interpretations, and the post-interpretation survey.

Pathologists were offered up to 20 hours of free category 1 Continuing Medical Education credits for completing the baseline survey, the interpretations, a web based educational feedback module, and a brief post-interpretation survey. In the post-interpretation survey, we asked participants whether standardized classification for melanocytic lesions would improve patient care, and their likelihood of adopting it in practice.

Statistical analyses

We calculated both interobserver and intraobserver concordance rates. For interobserver concordance, we considered all pairs of interpretations of the same case by two different pathologists in phase 1 and calculated the proportion of those pairs where interpretations were in the same diagnostic class. The 187 pathologists in phase 1 provided 6814 participant pairs reading the same test set and 48 case interpretations per pair, resulting in 327 072 total pairwise assessments of two pathologists independently interpreting the same glass slide for a case. Confidence intervals for interobserver rates used centiles of the bootstrap distribution, where resampling of pathologists was performed 3000 times.

For intraobserver concordance, we calculated the proportion of phase 1 interpretations where the same slide received an interpretation in the same diagnostic class by the same pathologist in phase 2. Confidence intervals for intraobserver concordance rates used a logit transformation and robust standard error that accounted for clustering at pathologist level. As a precaution, we created back-up test sets utilizing adjacent serial sections reviewed by the consensus panel for comparability. One of the five test sets was lost in shipment during phase 2. The intraobserver phase 1 or 2 reproducibility for the 11 participants who interpreted the back-up test set in phase 2 was similar to that of the 10 participants reading the original set in both phases. Change in overall reproducibility when excluding the former was negligible (3418/5112, 66.9% to 3116/4644, 67.1%).

To measure accuracy, we compared the pathologists' diagnoses on each case with one of three reference diagnoses. While our primary reference diagnosis was the consensus reference diagnosis reached by the aforementioned panel of three experienced dermatopathologists, two additional reference diagnoses were explored. We defined an experienced participant reference diagnosis based on the most frequent classification (mode) of each case by the board certified and/or dermatopathology fellowship trained participants (74 of the 187 pathologists completing phase 1). For 12 cases, a bimodal distribution was observed, and we chose the more severe diagnosis. The third reference standard was the community reference diagnosis defined by the most frequent diagnosis (mode) of each case by all participating pathologists. In three of the 240 patient cases, a bimodal distribution occurred in the community reference diagnosis, and we chose the more severe diagnosis.

Primary accuracy outcome measures included over-interpretation, under-interpretation, and overall concordance rates with the reference diagnoses. We defined over-interpretation as diagnosing cases at a higher diagnostic class than the reference diagnosis, and under-interpretation as diagnosing cases at a lower diagnostic class than the reference diagnosis. Interpretations in agreement were those in which the diagnostic classes assigned by the participants and the reference diagnoses were concordant. Confidence intervals accounted for both within participant and across participant variability. We also investigated variability in participant interpretations of each case separately to assess whether variability was limited to a subset of cases.

Population estimates

We estimated the probability that a pathologist's interpretation of a skin biopsy slide at the US population level would be confirmed if reviewed by a consensus based reference standard derived from three experienced dermatopathologists interpreting the same slide. For example, if one slide from a patient's skin biopsy is initially interpreted as melanoma in situ (MPATH-Dx class III), how likely is this patient to obtain the same diagnosis if a panel of three experienced pathologists provides a consensus interpretation of the same slide?

This calculation required reweighting the prevalence of skin biopsy classifications to reflect the distribution found in clinical practice compared with the distribution in our study, which included more of the cases that were intermediate and more difficult to interpret. Recent results about the prevalence of skin pathology diagnoses of melanocytic lesions from a large health plan in the Pacific Northwest of the USA (J P Lott, personal communication, 2017) were employed, where the prevalence values were 83.1% (15 558/18 715) for class I, 8.3% (1548/18 715) for class II, 4.5% (842/18 715) for class III, 2.2% (405/18 715) for class IV, and 1.9% (362/18 715) for class V (see table 3 in web appendix). In comparison, the prevalence values in our study were 10% (25/240) for class I, 15% (36/240) for class II, 25% (60/240) for class III, 24% (58/240) for class IV, and 25% (61/240) for class V (see table 2 in web appendix). The method for calculating the probabilities using Bayes' theorem have previously been described[19] and is a standard algorithm for the calculation of predictive values from accuracy and prevalence estimates. The method essentially involves reweighting case interpretations by the ratio of the population prevalence to the study prevalence of the diagnostic category assigned to that case by the reference.

Patient involvement

This work was inspired by the first author's experience as a patient undergoing a skin biopsy, which resulted in three different independent interpretations, ranging from benign to invasive melanoma. No other patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for recruitment, design, or implementation of the study. No patients were asked to advise on interpretation or writing up of results. We look forward to collaborating with both patient and professional groups to disseminate our findings broadly, with the goal of increasing understanding of variability in diagnostic interpretation and ultimately to improve patient care.

Results

Pathologist participation and characteristics

Of 301 eligible participants, 207 (69%) were enrolled and 187 (62%) completed independent phase 1 interpretations. There were no statistically significant differences in mean age, time spent in direct medical care, or number practicing in a community of 250 000 or more people between the 207 eligible pathologists who agreed to participate and the pathologists who were eligible but declined (data not shown, all comparisons P≥0.05). Among eligible responders, a slightly higher percentage of women (84/111, 76%) than men (123/190, 65%; P=0.048) participated.

Of those completing phase 1, 99 participants agreed to participate in the aforementioned substudy of digital whole slide imaging in phase 2, and were randomized to interpret glass (n=49) or digital (n=50) subsets of 36 cases (see figure 1 in web appendix). Those who declined the substudy (n=74) received their same set of 48 glass slides for phase 2 interpretations. A total of 118 participants completed phase 2 in the glass format and were retained for intraobserver analyses.

Table 2 shows the characteristics of the 187 participating pathologists. Most were men (n=114, 61%), aged 50 or more years (n=100, 54%), not affiliated with an academic medical center (n=134, 72%), and reported 10 or more years of experience in interpreting melanocytic skin lesions (n=113, 60%). All pathologists interpreted melanocytic skin lesions in their clinical practice as a requirement to participate; for 36 (19%) these melanocytic lesions represented more than a quarter of their caseload. Though 129 (69%) reported that interpreting melanocytic skin lesions made them more nervous than other types of pathology, 161 (86%) also reported being moderately to extremely confident in their assessments of melanocytic lesions.

Diagnostic terms for melanocytic lesions

The 240 biopsy cases were divided into five test sets (A to E). Each test set in phase 1 had 48 cases and each test set in phase 2 had 36 or 48 cases, as previously described. The 187 pathologists were randomized to a test set, with the final number of pathologists interpreting each test set in phase 1: test sets A, n=39; B, n=36; C, n=38; D, n=36; and E, n=38. Thus 36 to 39 different pathologists in phase 1 independently interpreted the same original glass slide for each skin biopsy case, with each pathologist viewing the same glass slide when interpreting a case.

The pathologists used diverse diagnostic terms to classify the melanocytic proliferations. The mean number of diagnostic terms applied for each case in phase 1 was 10 (SD 4, range 2-21). For example, one case independently interpreted by 36 study pathologists using their own microscopes to view the same glass slide had 18 different terms ascribed to it for the diagnosis (fig 1). Despite the striking variation in terminology, the suggested treatment for many of these diagnostic labels using Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx) would be the same, highlighting the importance of the MPATH-Dx classification tool to organize extant non-standardized diagnostic terminology into a smaller number of simpler categories, which range from class I (eg, benign melanocytic lesions) to class V (≥pT1b invasive melanoma).

Reproducibility of diagnoses

Intraobserver data were assessed for 118 pathologists based on phase 1 and 2 interpretations of the same cases at least eight months apart (table 3). Cases interpreted in phase 1 as class I (eg, banal or mildly dysplastic nevus) and class V (eg, ≥pT1b invasive melanoma) were likely to receive a diagnosis in the same class when interpreted by the same pathologist in phase 2 (77% (1155/1506) and 83% (852/1031), respectively). Pathologists' reproducibility when diagnosing the same case twice was lower for cases initially interpreted as class II (eg, moderately dysplastic nevus; 35% (227/644)), class III (eg, melanoma in situ; 60% (653/1091)), and class IV (eg, early stage invasive melanoma; 63% (531/840)).

Table 2 | Self reported characteristics of participating pathologists who completed the baseline survey and phase 1 interpretations (n=187)

Physician characteristics	No (%)
Demographics	
Age (years):	
<40	31 (17)
40-49	56 (30)
50-59	63 (34)
≥60	37 (20)
Sex:	
Female	73 (39)
Male	114 (61)
Training and experience	
Affiliation with academic medical centre:	
No	134 (72)
Yes, adjunct or affiliated	34 (18)
Yes, primary appointment	19 (10)
Residency specialty:	
Anatomic or clinical pathology	168 (90)
Dermatology	15 (8)
Both dermatology and anatomic or clinical pathology	4 (2)
Training:	
Board certified or fellowship trained in dermatopathology*	74 (40)
Other board certification or fellowship training†	113 (60)
Years interpreting melanocytic skin lesions:	
<5	29 (16)
5-9	45 (24)
10-19	57 (30)
≥20	56 (30)
Per cent of caseload interpreting melanocytic skin lesions:	
<10	79 (42)
10-24	72 (38)
25-49	28 (15)
≥50	8 (4)
Average No of melanoma cases (melanoma in situ and invasive melanoma) interpreted each month:	
<5	82 (44)
5-9	47 (25)
≥10	58 (31)
Average No of benign melanocytic skin lesions interpreted each month:	
<25	54 (29)
25-49	32 (17)
50-149	51 (27)
≥150	50 (27)
Considered an expert in melanocytic skin lesions by colleagues:	
No	108 (58)
Yes	79 (42)
Feelings and thoughts about interpreting melanocytic skin lesions	
In general, how challenging do you find interpreting melanocytic skin lesions?:	
Challenging (somewhat challenging to very challenging)	179 (96)
Easy (very easy to somewhat easy)	8 (4)
Interpreting melanocytic skin lesions makes me more nervous than other types of pathology:	
Agree (slightly agree to strongly agree)	129 (69)
Disagree (strongly disagree to slightly disagree)	58 (31)
How confident are you in your assessments of melanocytic skin lesions?:	
Confident (extremely confident to moderately confident)	161 (86)
Not confident (somewhat confident to not at all confident)	26 (14)

*Consists of physicians with single or multiple fellowships that include dermatopathology, and physicians with single or multiple board certifications that include dermatopathology.
†Includes fellowships or board certifications in surgical pathology, cytopathology, or hematopathology.

As expected, pathologists were more consistent with their own initial diagnosis of a case when viewing a glass slide a second time than when their diagnoses were compared with peers independently interpreting the same glass slide. Intraobserver concordance rates were consistently higher than the interobserver concordance rates (table 4). For example, the intraobserver and interobserver concordance rates for cases interpreted in phase 1 as class IV were 63% (95% confidence interval 59% to 67%) and 46% (43% to 49%), respectively.

Accuracy of diagnoses

Table 5 and figure 2 show the accuracy of the 187 pathologists' phase 1 interpretations for each diagnostic class based on the consensus reference diagnosis. Concordance with the reference was 92% (862/935) for class I, 25% (331/1342) for class II, 40% (908/2247) for class III, 43% (928/2169) for class IV, and 72% (1646/2283) for class V. Figure 3 shows a comparison of the pathologists' over-interpretation and under-interpretation rates when considering the consensus reference panel and the two additional reference diagnoses. The discordance rates were more than 40% for cases in classes II, III, and IV regardless of the method used to define the reference standard.

Population estimates

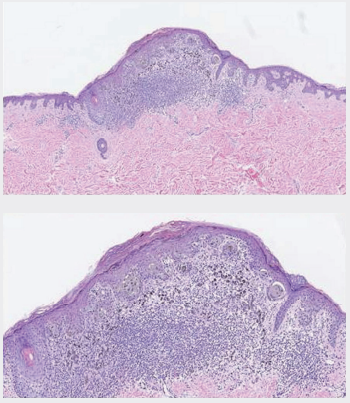
Most melanocytic skin biopsy lesions in a clinical setting are of benign MPATH-Dx class I lesions, whereas the composition of our test set was heavily weighted towards MPATH-Dx classes II-V lesions. We describe at a population perspective how the diagnostic variability noted in this testing situation might affect accuracy (fig 4, table 3 in web appendix). Overall, 82.8% (95% confidence interval 81.0% to 84.5%) of skin biopsy diagnoses for melanocytic lesions would be verified by consensus of three experienced dermatopathologists, with 8.0% (6.2% to 9.9%) of biopsies over-interpreted by the initial pathologist and 9.2% (8.8% to 9.6%) under-interpreted. Of cases interpreted in classes II, III, and IV, we estimate that only 26%, 35%, and 51%, respectively would be confirmed by the consensus reference diagnosis, whereas the numbers are 92% for class I and 78% for class V (see table 3 in web appendix). Of patients classified in categories IV or V (eg, with invasive melanoma) by study pathologists we estimate from table 3 in the web appendix that 16% (52.4/324.1) would be reclassified downward to benign categories I or II by the experienced consensus panel. Of patients classified in categories I or II by study pathologists we estimate that 0.5% (41.3/8186.9) would be classified in categories IV or V by the experienced consensus panel.

Post-interpretation survey

Most pathologists (96%) thought it somewhat to very likely that patient care would be improved by the use of a standardized taxonomy such as the MPATH-Dx tool in the diagnosis of melanocytic skin lesions. Nearly all participants (98%) also stated that they would likely adopt a standardized taxonomy in their own clinical practice if available.

Discussion

This study highlights challenges and also limitations in the diagnosis of melanocytic skin lesions by current



Diagnostic terms given	No of pathologists
MPATH-Dx class I	
Common nevus, junctional	3
Dysplastic nevus - mild	2
Halo nevus (1)	1
Atypical melanocytic neoplasm, junctional (suggested treatment of no further treatment required)	1
MPATH-Dx class II	
Spitz nevus (conventional), (junctional, compound, or intradermal)	4
Dysplastic nevus - moderate	2
Pigmented spindle cell nevus (junctional or compound)	1
Atypical nevus not otherwise specified, including atypical nevus of special anatomic - moderate	1
Atypical intraepithelial melanocytic proliferation (AIMP) (suggested treatment of repeat excision <5 mm margins (narrow but complete))	1
Atypical melanocytic neoplasm, junctional (suggested treatment of repeat excision <5 mm margins (narrow but complete))	1
MPATH-Dx class III	
Atypical/dysplastic Spitz lesion, (junctional, compound, or dermal)	5
Melanoma in situ, common/pagetoid/superfical spreading	5
Dysplastic nevus - severe	1
Atypical nevus not otherwise specified, including atypical nevus of special anatomic site - severe	1
Melanoma in situ, not otherwise specified	1
Atypical melanocytic neoplasm, junctional (suggested treatment of repeat excision with at least 5mm (but <1 cm) margins)	1
MPATH-Dx class IV	
Invasive melanoma, superficial spreading melanoma	4
MPATH-Dx class V	
Invasive melanoma, heavily pigmented melanoma	1
Total	**36**

Fig 1 | Diagnostic terms given to example case by 36 pathologists who each independently interpreted the same glass slide (top image 5× magnification, bottom image 10× magnification)

Table 3 | Intraobserver concordance of 118 pathologists' interpretations of melanocytic skin biopsy lesions of the same case at phase 1 and phase 2 at least eight months apart*

Phase 1 diagnosis	Phase 2 diagnosis (No of paired interpretations)						Intraobserver concordance† % (95% CI)
	Class I	Class II	Class III	Class IV	Class V	Total	
Class I	**1155**	188	119	27	17	**1506**	77 (73 to 80)
Class II	170	**227**	182	37	28	**644**	35 (31 to 39)
Class III	91	120	**653**	169	58	**1091**	60 (56 to 63)
Class IV	20	37	147	**531**	105	**840**	63 (59 to 67)
Class V	14	16	44	105	**852**	**1031**	83 (80 to 85)
Total	**1450**	**588**	**1145**	**869**	**1060**	**5112**	67‡ (65 to 69)

Numbers of diagnostic interpretations with intraobserver agreement are emboldened.
*Concordance rates are influenced by case composition, which included larger proportions of cases in classes II-V than would typically be encountered in practice.
†Denominator is phase 1 interpretations and numerator is phase 2 assessments that agreed with phase 1 interpretation. Does not include participants who reviewed glass slides in phase 1 and digital images in phase 2.
‡Average κ for intraobserver agreement across participants is 0.57.

Table 4 | Interobserver concordance of pathologists' interpretations of melanocytic skin biopsy lesions. Pairwise comparison of interpretations by 187 participating pathologists in phase 1. Diagnoses for all possible ordered pairs of participants reading the same glass slide are included*

First pathologist's interpretation	Second pathologist's interpretation						Interobserver concordance % (95% CI)
	Class I	Class II	Class III	Class IV	Class V	Total	
Class I	**64 122**	15 082	11 223	2993	1773	**95 193**	71 (69 to 73)
Class II	15 082	**10 366**	11 028	3974	1903	**42 353**	25 (22 to 27)
Class III	11 223	11 028	**31 326**	12 675	4494	**70 746**	45 (42 to 47)
Class IV	2993	3974	12 675	**22 334**	8854	**50 830**	46 (43 to 49)
Class V	1773	1903	4494	8854	**50 926**	**67 950**	77 (75 to 79)
Total	**95 193**	**42 353**	**70 746**	**50 830**	**67 950**	**327 072**	55† (53 to 56)

Numbers of interpretations with agreement are emboldened.
*Average pairwise agreement is an unweighted average across all participant pairs. The number of first pathologist interpretations for a given diagnostic class varies across pairs. Concordance rates are influenced by case composition, which included larger proportions of cases in classes II-V than would typically be encountered in practice. There are 6814 distinct order participant pairs ×48 case interpretations per pair yielding 327 072 interpretations.
†Average κ for interobserver agreement across all participant pairs is 0.42.

practicing pathologists. The highest levels of accuracy were attained for class I benign nevi (92%) and class V high stage invasive melanoma (72%); these cases are at the polar ends of the histopathologic spectrum. Interpretations for cases in the middle of the spectrum had noticeably lower accuracy, as less than 50% of the diagnoses were in concordance with the reference diagnoses; class II, moderate dysplastic nevus (25%); class III, melanoma in situ (40%); and class IV, early stage invasive melanoma (43%). Pathologists' interpretations of the same case on two occasions also lacked reproducibility for cases in the middle of the spectrum. This low level of diagnostic precision is of clinical concern. Although diagnostic discordance has been described in other areas of clinical medicine,[20 21] including pathologists diagnosing breast biopsies[22] and radiologists interpreting mammograms,[23] the findings reported here are more pronounced than in other disciplines of medicine.

Comparison with other studies

The results of this large study strongly validate the conclusions from smaller studies that histological diagnoses of melanocytic lesions can vary among pathologists. Previous studies were constrained by small numbers of cases (eg, ≤20),[6 24-26] non-randomly selected cases,[27-37] narrow disease spectrum of cases,[5 7 8 25-27 29 32 35 36 38-45] smaller numbers of physicians (eg, ≤10),[6-8 24 25 27 30 34-42 44 46-48] or exclusive testing of expert pathologists.[5 7 8 16 30 34 35 37-42 47] In contrast, our study is the most extensive evaluation to date. We selected 240 cases from the full histopathologic spectrum and enrolled 187 practicing pathologists from diverse geographic locations and clinical settings. Unique to our study are estimates of intraobserver reproducibility, as well as accuracy defined by three different reference diagnoses and presentation of the impact of diagnostic variability at a US population perspective.

Table 5 | Accuracy of 187 participating pathologists' when phase 1 interpretations are compared with the consensus reference diagnoses*

Consensus reference diagnosis†	Study pathologists' interpretation					Total interpretations (No)	% Concordance with reference diagnosis (95% CI)
	Class I	Class II	Class III	Class IV	Class V		
Class I	**862**	50	19	3	1	935	92 (90 to 94)
Class II	843	**331**	131	26	11	1342	25 (22 to 28)
Class III	695	520	**908**	113	11	2247	40 (37 to 44)
Class IV	150	176	717	**928**	198	2169	43 (39 to 46)
Class V	68	87	161	321	**1646**	2283	72 (69 to 75)
Total	**2618**	**1164**	**1936**	**1391**	**1867**	8976	

*Concordance in interpretation is emboldened.
†Reference diagnosis was obtained from consensus of three experienced dermatopathologists.

Strengths and limitations of this study

In the absence of a biological reference standard for diagnosing melanocytic lesions, our analytic approach is strengthened by the use of three reference standards to estimate accuracy. Our reference standard based on the consensus of three experienced dermatopathologists would be considered ideal in clinical practice. In addition, we provided sensitivity analyses based on two other reference standards: the majority diagnosis of participating board certified or fellowship trained dermatopathologists, and a community reference based on the mode of all study pathologists. As some pathologists may have participated in the study to improve their own performance, the community reference may not be ideal. Although results differed by reference standard, accuracy was low for classes II, III, and IV cases, regardless of the reference standard used.

Evaluating diagnostic accuracy requires research methods and a study setting that deviates from normal clinical practice. While many skin biopsy cases have only one slide available and a pathologist's diagnosis often hinges on one area within a slide, in clinical practice pathologists might have the opportunity to review more slides on some of their patients. Pathologists might also be able to obtain second opinions from colleagues or request ancillary tests such as immunohistochemical and molecular studies before rendering a diagnosis. While our study evaluated pathologists' independent interpretations using their own microscopes, it did not evaluate overall processes within health systems. Ideally we would insert biopsy cases into the clinical practices of a diverse group of pathologists in a large scale, blinded manner, yet this design would not be logistically feasible given the large number of pathologists in our study from many diverse clinical practices and the high number of cases they each interpreted. The cases in our study also included more lesions in classes II-V relative to class I than is typical of clinical practice, thus the importance of our population perspective results.

Clinical and policy implications

Disease classification systems often evolve from examination of prototypical cases by experts in the specialty and are then disseminated to a broader range of practitioners on the full breadth of disease in clinical practice. Without validation of extant diagnostic criteria that are applied to the millions of skin biopsies performed annually, present diagnoses may not reliably or accurately distinguish biologically important differences.

Given the striking array of diagnostic terms used by practicing pathologists when interpreting the same melanocytic lesion, we recommend further studies of a simple, management oriented classification system, such as the Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx) system used in this study.[13] Most participating pathologists thought patient care would be improved through use of such a simplified taxonomy. Thus, reducing more than 50 terms currently used into a smaller number of classes may improve communication between the pathologist and the patient's primary clinician. With patients increasingly reading their own pathology reports, often through secure electronic portals, increased clarity is also important.[49 50] The MPATH-Dx classification system will likely require further examination and revisions, given the low levels of reproducibility and accuracy for cases in the middle of the histopathologic spectrum, and studies of education in the definitions of these subjective categories are needed.

We also recommend transparency and acknowledgement of the inherent limits of our ability to classify melanocytic lesions. Communicating the degree of diagnostic uncertainty is an important part of professional practice, with reports showing that 71% of consultations between clinicians and patients include verbal expressions of uncertainty.[51] Pathology reports often include phrases describing uncertainty of the diagnosis, yet interpretation of these phrases varies widely.[52] National guidelines on phraseology in pathology reporting have long been suggested.[53] We propose adding standardized statements to pathology reports reminding readers that melanocytic lesions are challenging to interpret and that variability exists among pathologists in interpretation, especially in the middle diagnostic classes. When similar evidence summary statements were added to radiology reports of spine imaging, fewer narcotics were prescribed by physicians receiving the reports.[54] The impact of adding such disclaimers on clinical care and also on malpractice lawsuits should be studied.

Finally, reliable and objective techniques need to be developed and validated to support pathologists' visual assessments of melanocytic lesions. We hope that future systems using digital whole slide imaging platforms to obtain second opinions or molecular analysis of skin biopsies can be developed and evaluated that may lead to more definitive classification of melanocytic lesions.[55 56]

Conclusion

Our study emphasizes persistent difficulties with classifying medical data based on subjective interpretations. In this large study, diagnoses of melanoma in situ and early stage invasive melanoma (pT1a according to the American Joint Committee on Cancer seventh edition cancer staging manual), together more common than

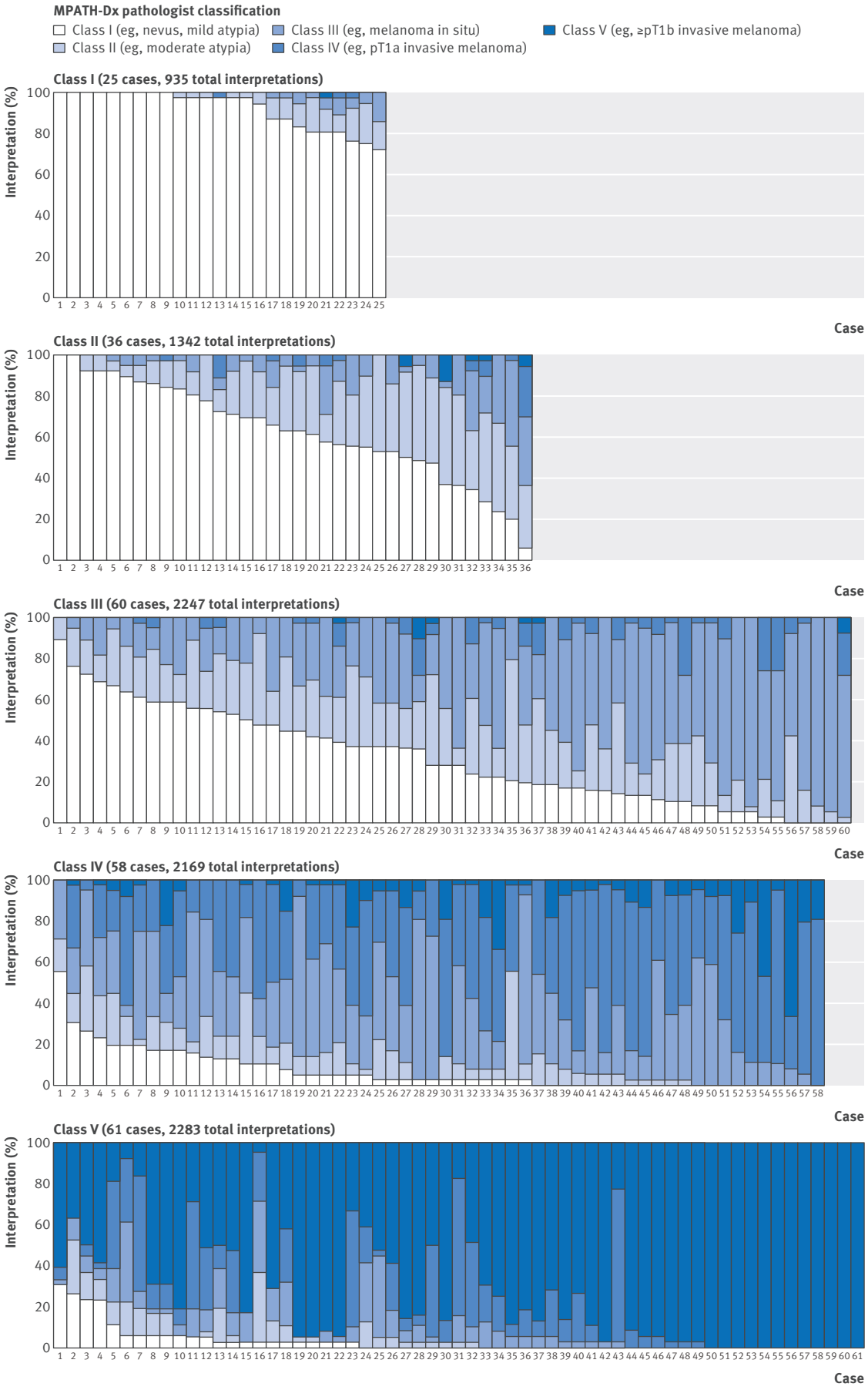


Fig 2 | Participant interpretive variation on each of 240 cases, with cases organized based on the MPATH-Dx consensus reference diagnosis class

doi: 10.1136/bmj.j2813 | *BMJ* 2017;357:j2813 | the bmj

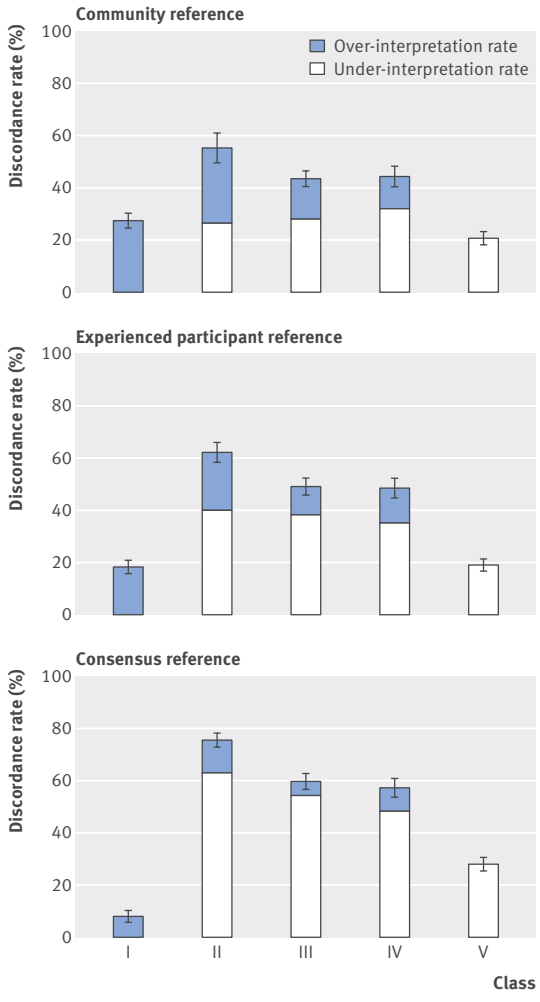


Fig 3 | Comparison of accuracy (discordance rates of over-interpretation rate and under-interpretation rate) by all three reference diagnoses

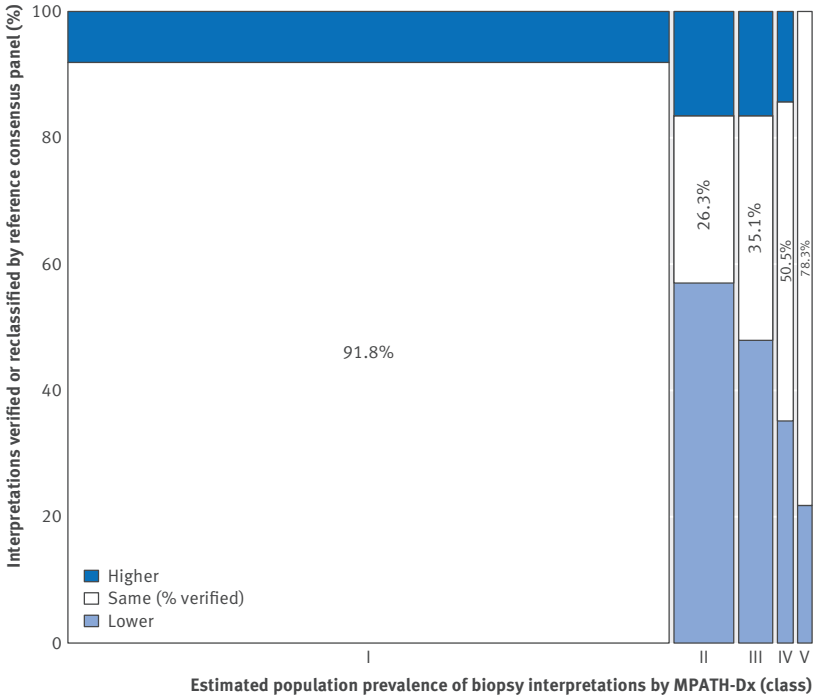


Fig 4 | Population level predicted proportions of cutaneous melanocytic biopsy interpretations that would be verified by the consensus reference panel or would be classified as over-interpretations or under-interpretations

all other stages of melanoma combined, were neither reproducible nor accurate. Efforts to improve clinical practice should include simplification of terminology by use of a standardized classification system, acknowledgment of the extant uncertainty of specific diagnoses in pathology reports, and development of more sophisticated diagnostic tools to support pathologists.

AUTHOR AFFILIATIONS

[1]Department of Medicine, University of Washington School of Medicine, Seattle, WA, 98104, USA
[2]Department of Pathology, Institut Curie Institute Hospital, University of Paris Descartes Faculty of Medicine University, Paris, France
[3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
[4]Program in Biostatistics, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[5]Department of Family Medicine, Oregon Health & Science University, Portland, OR, USA
[6]Departments of Epidemiology and Pediatrics, Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH, USA
[7]Departments of Medical Informatics and Clinical Epidemiology and Medicine, School of Medicine, Oregon Health & Science University, Portland, OR, USA
[8]Providence Cancer Center, Providence Health and Services, Portland, OR, USA
[9]Geisel School of Medicine at Dartmouth, Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, USA
[10]Department of Biomedical Data Science, Department of Epidemiology, Norris Cotton Cancer Center, Lebanon, NH, USA
[11]Departments of Medicine and Community and Family Medicine, The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH, USA
[12]Center for Dermatoepidemiology, Providence VA Medical Center, Providence, RI, USA
[13]Departments of Dermatology and Epidemiology, Brown University, Providence, RI, USA
[14]Pathology Associates, Clovis, CA, USA
[15]Division of Dermatology, Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA
[16]Dermatopathology Northwest, Bellevue, WA, USA

The American Medical Association master file is the source for some of the data used in comparing characteristics of participants and non-participants. We thank the study participants for their commitment to improving clinical care in dermatopathology.

Contributors: JGE, GML, and MSP had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. JGE, LMR, GML, and MSP acquired, analyzed, and interpreted the data. JGE, RLB, DEE, and MWP drafted the manuscript. GML and MSP performed the statistical analysis. JGE obtained the funding. LMR provided administrative, technical, and material support. JGE and LMR supervised the study. All authors contributed to the overall conception and design of the study and revised the manuscript for intellectual content. JGE is the guarantor.

Funding: This work was supported by the National Cancer Institute (R01 CA151306, R01 CA201376 and K05 CA104699). The funding agency had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf. All authors had financial support from the National Cancer Institute for the submitted work. RLB reports a financial relationship with Myriad Genetics outside of the submitted work, and GML reports grants from Fred Hutchinson Cancer Research Center during the conduct of the study.

Ethical approval: This study was approved by the institutional review boards of Dartmouth College (No 22983), the Fred Hutchinson Cancer Research Center (No 7573), Providence Health and Services of Oregon (No 00242), and the University of Washington (No 44309). All participating pathologists signed an informed consent form.

Informed consent was not required of the patients whose biopsy specimens were included.

Data sharing: Details on how to obtain additional data from the study (eg, statistical code, datasets) are available from the corresponding author.

Transparency: The lead author (JGE) affirms that this manuscript is an honest, accurate, and transparent account of the study being reported, that no important aspects of the study have been omitted, and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 *National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care*.The National Academies Press, 2015.
2 Feinstein AR. Boolean Algebra and Clinical Taxonomy. *N Engl J Med* 1963;269:929-38. doi:10.1056/NEJM196310312691801.
3 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926-30. doi:10.1056/NEJM197810262991705.
4 Welch HG, Woloshin S, Schwartz LM. Skin biopsy rates and incidence of melanoma: population based ecological study. *BMJ* 2005;331:481. doi:10.1136/bmj.38516.649537.E0.
5 Gerami P, Busam K, Cochran A, et al. Histomorphologic assessment and interobserver diagnostic reproducibility of atypical spitzoid melanocytic neoplasms with long-term follow-up. *Am J Surg Pathol* 2014;38:934-40. doi:10.1097/PAS.0000000000000198.
6 Duncan LM, Berwick M, Bruijn JA, Byers HR, Mihm MC, Barnhill RL. Histopathologic recognition and grading of dysplastic melanocytic nevi: an interobserver agreement study. *J Invest Dermatol* 1993;100:318S-21S. doi:10.1038/jid.1993.55.
7 Duray PH, DerSimonian R, Barnhill R, et al. An analysis of interobserver recognition of the histopathologic features of dysplastic nevi from a mixed group of nevomelanocytic lesions. *J Am Acad Dermatol* 1992;27:741-9. doi:10.1016/0190-9622(92)70248-E.
8 Corona R, Mele A, Amini M, et al. Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. *J Clin Oncol* 1996;14:1218-23. doi:10.1200/JCO.1996.14.4.1218.
9 Swerlick RA, Chen S. The melanoma epidemic. Is increased surveillance the solution or the problem?*Arch Dermatol* 1996;132:881-4. doi:10.1001/archderm.1996.03890320029004.
10 Onega T, Reisch LM, Frederick PD, et al. Use of Digital Whole Slide Imaging in Dermatopathology. *J Digit Imaging* 2016;29:243-53. doi:10.1007/s10278-015-9836-y.
11 Knezevich SR, Barnhill RL, Elder DE, et al. Variability in mitotic figures in serial sections of thin melanomas. *J Am Acad Dermatol* 2014;71:1204-11. doi:10.1016/j.jaad.2014.07.056.
12 Carney PA, Frederick PD, Reisch LM, et al. How concerns and experiences with medical malpractice affect dermatopathologists' perceptions of their diagnostic practices when interpreting cutaneous melanocytic lesions. *J Am Acad Dermatol* 2016;74:317-24, quiz 324.e1-8.
13 Piepkorn MW, Barnhill RL, Elder DE, et al. The MPATH-Dx reporting schema for melanocytic proliferations and melanoma. *J Am Acad Dermatol* 2014;70:131-41. doi:10.1016/j.jaad.2013.07.027.
14 Carney PA, Reisch LM, Piepkorn MW, et al. Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified Delphi method. *J Cutan Pathol* 2016;43:830-7. doi:10.1111/cup.12751.
15 Edge SB, Byrd DR, *Compton CC, Fritz AG, Greene FL, Trotti A, eds. AJCC cancer staging manual (7th ed)*. Springer; 2010.
16 Lott JP, Elmore JG, Zhao GA, et al. International Melanoma Pathology Study Group. Evaluation of the Melanocytic Pathology Assessment Tool and Hierarchy for Diagnosis (MPATH-Dx) classification scheme for diagnosis of cutaneous melanocytic neoplasms: Results from the International Melanoma Pathology Study Group. *J Am Acad Dermatol* 2016;75:356-63. doi:10.1016/j.jaad.2016.04.052.
17 Dalkey NC, Brown B, Cochran N. *The Delphi Method, III. Use of Self Ratings to Improve Group Estimates*.Rand Corp, 1969.
18 American Medical Association. Physicians. Secondary Physicians 2011. http://www.dmddata.com/data_lists_physicians.asp.
19 Elmore JG, Nelson HD, Pepe MS, et al. Variability in Pathologists' Interpretations of Individual Breast Biopsy Slides: A Population Perspective. *Ann Intern Med* 2016;164:649-55. doi:10.7326/M15-0964.
20 Feinstein AR. A bibliography of publications on observer variability. *J Chronic Dis* 1985;38:619-32. doi:10.1016/0021-9681(85)90016-5.
21 Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). *J Clin Epidemiol* 1992;45:567-80. doi:10.1016/0895-4356(92)90128-A.
22 Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* 2015;313:1122-32. doi:10.1001/jama.2015.1405.
23 Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994;331:1493-9. doi:10.1056/NEJM199412013312206.
24 Carli P, De Giorgi V, Naldi L, Dosi G. Dermoscopy Panel. Reliability and inter-observer agreement of dermoscopic diagnosis of melanoma and melanocytic naevi. *Eur J Cancer Prev* 1998;7:397-402. doi:10.1097/00008469-199810000-00005.
25 Krieger N, Hiatt RA, Sagebiel RW, Clark WH Jr,, Mihm MC Jr. Inter-observer variability among pathologists' evaluation of malignant melanoma: effects upon an analytic study. *J Clin Epidemiol* 1994;47:897-902. doi:10.1016/0895-4356(94)90193-7.
26 CRC Melanoma Pathology Panel. A nationwide survey of observer variation in the diagnosis of thin cutaneous malignant melanoma including the MIN terminology. *J Clin Pathol* 1997;50:202-5. doi:10.1136/jcp.50.3.202.
27 Heenan PJ, Matz LR, Blackwell JB, et al. Inter-observer variation between pathologists in the classification of cutaneous malignant melanoma in western Australia. *Histopathology* 1984;8:717-29. doi:10.1111/j.1365-2559.1984.tb02388.x.
28 Murali R, Hughes MT, Fitzgerald P, Thompson JF, Scolyer RA. Interobserver variation in the histopathologic reporting of key prognostic parameters, particularly clark level, affects pathologic staging of primary cutaneous melanoma. *Ann Surg* 2009;249:641-7. doi:10.1097/SLA.0b013e31819ed973.
29 Shoo BA, Sagebiel RW, Kashani-Sabet M. Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center. *J Am Acad Dermatol* 2010;62:751-6. doi:10.1016/j.jaad.2009.09.043.
30 Braun RP, Gutkowicz-Krusin D, Rabinovitz H, et al. Agreement of dermatopathologists in the evaluation of clinically difficult melanocytic lesions: how golden is the 'gold standard'?*Dermatology* 2012;224:51-8. doi:10.1159/000336886.
31 Gaudi S, Zarandona JM, Raab SS, English JC 3rd,, Jukic DM. Discrepancies in dermatopathology diagnoses: the role of second review policies and dermatopathology fellowship training. *J Am Acad Dermatol* 2013;68:119-28. doi:10.1016/j.jaad.2012.06.034.
32 Patrawala S, Maley A, Greskovich C, et al. Discordance of histopathologic parameters in cutaneous melanoma: Clinical implications. *J Am Acad Dermatol* 2016;74:75-80. doi:10.1016/j.jaad.2015.09.008.
33 Niebling MG, Haydu LE, Karim RZ, Thompson JF, Scolyer RA. Pathology review significantly affects diagnosis and treatment of melanoma patients: an analysis of 5011 patients treated at a melanoma treatment center. *Ann Surg Oncol* 2014;21:2245-51. doi:10.1245/s10434-014-3682-x.
34 Farmer ER, Gonin R, Hanna MP. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. *Hum Pathol* 1996;27:528-31. doi:10.1016/S0046-8177(96)90157-4.
35 Barnhill RL, Argenyi ZB, From L, et al. Atypical Spitz nevi/tumors: lack of consensus for diagnosis, discrimination from melanoma, and prediction of outcome. *Hum Pathol* 1999;30:513-20. doi:10.1016/S0046-8177(99)90193-4.
36 Meyer LJ, Piepkorn M, Goldgar DE, et al. Interobserver concordance in discriminating clinical atypia of melanocytic nevi, and correlations with histologic atypia. *J Am Acad Dermatol* 1996;34:618-25. doi:10.1016/S0190-9622(96)80061-2.
37 Ferrara G, Argenziano G, Soyer HP, et al. Dermoscopic and histopathologic diagnosis of equivocal melanocytic skin lesions: an interdisciplinary study on 107 cases. *Cancer* 2002;95:1094-100. doi:10.1002/cncr.10768.
38 Colloby PS, West KP, Fletcher A. Observer variation in the measurement of Breslow depth and Clark's level in thin cutaneous malignant melanoma. *J Pathol* 1991;163:245-50. doi:10.1002/path.1711630310.
39 Cook MG, Clarke TJ, Humphreys S, et al. The evaluation of diagnostic and prognostic criteria and the terminology of thin cutaneous malignant melanoma by the CRC Melanoma Pathology Panel. *Histopathology* 1996;28:497-512. doi:10.1046/j.1365-2559.1996.d01-464.x.
40 Lock-Andersen J, Hou-Jensen K, Hansen JP, Jensen NK, Søgaard H, Andersen PK. Observer variation in histological classification of cutaneous malignant melanoma. *Scand J Plast Reconstr Surg Hand Surg* 1995;29:141-8. doi:10.3109/02844319509034330.
41 Spatz A, Cook MG, Elder DE, Piepkorn M, Ruiter DJ, Barnhill RL. Interobserver reproducibility of ulceration assessment in primary cutaneous melanomas. *Eur J Cancer* 2003;39:1861-5. doi:10.1016/S0959-8049(03)00325-3.

42 Wechsler J, Bastuji-Garin S, Spatz A, et al. French Cutaneous Cancerology Group. Reliability of the histopathologic diagnosis of malignant melanoma in childhood. *Arch Dermatol* 2002;138:625-8. doi:10.1001/archderm.138.5.625.
43 Eriksson H, Frohm-Nilsson M, Hedblad MA, et al. Interobserver variability of histopathological prognostic parameters in cutaneous malignant melanoma: impact on patient management. *Acta Derm Venereol* 2013;93:411-6. doi:10.2340/00015555-1517.
44 Scolyer RA, Shaw HM, Thompson JF, et al. Interobserver reproducibility of histopathologic prognostic variables in primary cutaneous melanomas. *Am J Surg Pathol* 2003;27:1571-6. doi:10.1097/00000478-200312000-00011.
45 Brochez L, Verhaeghe E, Grosshans E, et al. Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions. *J Pathol* 2002;196:459-66. doi:10.1002/path.1061.
46 Boiko PE, Piepkorn MW. Reliability of skin biopsy pathology. *J Am Board Fam Pract* 1994;7:371-4.
47 Piepkorn MW, Barnhill RL, Cannon-Albright LA, et al. A multiobserver, population-based analysis of histologic dysplasia in melanocytic nevi. *J Am Acad Dermatol* 1994;30:707-14. doi:10.1016/S0190-9622(08)81499-5.
48 Weinstock MA, Barnhill RL, Rhodes AR, Brodsky GL. The Dysplastic Nevus Panel. Reliability of the histopathologic diagnosis of melanocytic dysplasia. *Arch Dermatol* 1997;133:953-8. doi:10.1001/archderm.1997.03890440019002.
49 Lott JP, Piepkorn MW, Elmore JG. Dermatology in an age of fully transparent electronic medical records. *JAMA Dermatol* 2015;151:477-8. doi:10.1001/jamadermatol.2014.4362.
50 Delbanco T, Walker J, Bell SK, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Ann Intern Med* 2012;157:461-70. doi:10.7326/0003-4819-157-7-201210020-00002.
51 Gordon GH, Joos SK, Byrne J. Physician expressions of uncertainty during patient encounters. *Patient Educ Couns* 2000;40:59-65. doi:10.1016/S0738-3991(99)00069-5.
52 Galloway M, Taiyeb T. The interpretation of phrases used to describe uncertainty in pathology reports. *Patholog Res Int* 2011;2011:656079.
53 Attanoos RL, Bull AD, Douglas-Jones AG, Fligelstone LJ, Semararo D. Phraseology in pathology reports. A comparative study of interpretation among pathologists and surgeons. *J Clin Pathol* 1996;49:79-81. doi:10.1136/jcp.49.1.79.
54 McCullough BJ, Johnson GR, Martin BI, Jarvik JG. Lumbar MR imaging and reporting epidemiology: do epidemiologic data in reports affect clinical management?*Radiology* 2012;262:941-6. doi:10.1148/radiol.11110618.
55 Shain AH, Yeh I, Kovalyshyn I, et al. The Genetic Evolution of Melanoma from Precursor Lesions. *N Engl J Med* 2015;373:1926-36. doi:10.1056/NEJMoa1502583.
56 Shain AH, Bastian BC. From melanocytes to melanomas. *Nat Rev Cancer* 2016;16:345-58. doi:10.1038/nrc.2016.37.

Appendix: Supplementary materials