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,1,2,11 Stevan R Knezevich,14 Michael W Piepkorn15,16

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.

Introduction
Diagnostic errors contribute to approximately 10% of patient deaths and are the top cause of medical malpractice payouts. 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.”

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. 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, 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,6-8 alluding to the potential for both overtreatment 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.11-16 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.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 |
|-----------------|--------------------------------|-------------------------|----------|
| O               | 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, pT1aT1) |
| 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 case-load 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 terms31 (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 participating pathologists. In three of the 240 patient 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 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. 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 board certified panel of three experienced dermatopathologists.

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 described19 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 2-22). 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)).
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% (41.3/8166.9) 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
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 1050                      | 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 interpretations.
‡Denominator is phase 1 interpretations and numerator is phase 2 assessments that agreed with phase 1 interpretations.

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                           | 64122 15082 11223 2993 1773 105933 | 71 (69 to 73)            |            |
| Class II                          | 15082 10366 11028 3976 1903 42353 | 25 (22 to 27)            |            |
| Class III                         | 11223 11028 31326 12675 4469 70746 | 45 (42 to 47)            |            |
| Class IV                          | 2993 3976 12675 22334 8854 50830 | 46 (43 to 49)            |            |
| Class V                           | 1773 1903 4494 8854 50926 67950 | 77 (75 to 79)            |            |
| Total                             | 95193 42353 70746 50830 67950 327072 | 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 x48 case interpretations per pair yielding 32072 interpretations.
†Average x for interobserver agreement across all participant pairs is 0.42.

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, s20,24-26 non-randomly selected cases,27-37 narrow disease spectrum of cases,5,7 8 25 27 39 36 38-45 smaller numbers of pathologists diagnosing breast biopsies28 and radiologists interpreting mammograms,29 the findings reported here are more pronounced than in other disciplines of medicine.
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                      | 862                               | 935                       | 92 (90 to 94)                                 |
| Class II                     | 843                               | 176                       | 71 (69 to 75)                                 |
| Class III                    | 695                               | 113                       | 40 (37 to 44)                                 |
| Class IV                     | 150                               | 198                       | 43 (39 to 46)                                 |
| Class V                      | 68                                | 1646                      | 72 (69 to 75)                                 |
| Total                        | 2618                              | 1391                      | 89%                                           |

*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.15 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.46 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 acknowledge- ment 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

**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
### Participant Interpretive Variation on Each of 240 Cases

Cases were organized based on the MPATH-Dx consensus reference diagnosis class.

#### Class I (25 cases, 935 total interpretations)

| Class I (eg, nevus, mild atypia) | Class III (eg, melanoma in situ) |
|-----------------------------------|----------------------------------|
| ![Interpretation Graph Class I](image) | ![Interpretation Graph Class III](image) |

#### Class II (36 cases, 1342 total interpretations)

| Class II (eg, moderate atypia) | Class IV (eg, pT1b invasive melanoma) |
|---------------------------------|--------------------------------------|
| ![Interpretation Graph Class II](image) | ![Interpretation Graph Class IV](image) |

#### Class III (60 cases, 2247 total interpretations)

| Class III (eg, melanoma in situ) | Class V (eg, ≥pT1b invasive melanoma) |
|----------------------------------|--------------------------------------|
| ![Interpretation Graph Class III](image) | ![Interpretation Graph Class V](image) |

#### Class IV (58 cases, 2169 total interpretations)

| Class IV (eg, pT1a invasive melanoma) | Class V (eg, ≥pT1b invasive melanoma) |
|--------------------------------------|--------------------------------------|
| ![Interpretation Graph Class IV](image) | ![Interpretation Graph Class V](image) |

#### Class V (61 cases, 2283 total interpretations)

| Class V (eg, ≥pT1b invasive melanoma) | Class V (eg, ≥pT1b invasive melanoma) |
|--------------------------------------|--------------------------------------|
| ![Interpretation Graph Class V](image) | ![Interpretation Graph Class V](image) |

*Fig 2 | Participant interpretive variation on each of 240 cases, with cases organized based on the MPATH-Dx consensus reference diagnosis class.*
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 extent uncertainty of specific diagnoses in pathology reports, and development of more sophisticated diagnostic tools to support pathologists.

**AUTHOR AFFILIATIONS**

1Department of Medicine, University of Washington School of Medicine, Seattle, WA, 98104, USA
2Department of Pathology, Institut Curie Institute Hospital, University of Paris Descartes Faculty of Medicine University, Paris, France
3Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
4Program in Biostatistics, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center Seattle, Center, Seattle, WA, USA
5Department of Family Medicine, Oregon Health & Science University, Portland, OR, USA
6Departments of Epidemiology and Pediatrics, Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH, USA
7Departments of Medical Informatics and Clinical Epidemiology and Medicine, School of Medicine, Oregon Health & Science University, Portland, OR, USA
8Providence Cancer Center, Providence Health Services, Portland, OR, USA
9Geisel School of Medicine at Dartmouth, Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, USA
10Department of Biomedical Data Science, Department of Epidemiology, Norris Cotton Cancer Center, Lebanon, NH, USA
11Departments 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
12Center for Dermatopepidemiology, Providence VA Medical Center, Providence, RI, USA
13Departments of Dermatology and Epidemiology, Brown University, Providence, RI, USA
14Pathology Associates, Clovis, CA, USA
15Division of Dermatology, Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA
16Dermatopathology 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 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 22993), 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 (e.g., statistical code, datasets) are available from the corresponding author.

Transparency: The lead author (IGE) 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.

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

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/NEJM196301262691801.
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/NEJM197810262991070.
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.38516.640957.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;38:914-40. doi:10.1097/PAS.0b013e3182a4300e.
6. Duncan LM, Benwick 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;100:315-21. doi:10.1093/jid/1993.55.
7. Duray PH, DerSimonian R, Barnhill RL, 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:10.1016/0900-9622(92)90248-E.
8. Conrad 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. 1978;299:926-30. doi:10.1056/NEJM197810262991070.
11. Faries PL, Heen RA, Chen S. The melanoma epidemic. Is increased surveillance the solution or the problem? Ann Intern Med 1994;121:1094-100. doi:10.1056/NEJM199403243311110.
12. Armer ER, Gonin R, Hanna MP. Discordance in the histopathologic diagnosis of melanoma in western Australia. Histopathology 1991;163:245-50. doi:10.1111/j.1365-2559.1984.tb02388.x.
13. Murali R, Hughes MT, Fitzgerald P, Thompson JF, Scoller RA. Interobserver variation in the histopathological 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.
14. Shoo BA, Sagebel RW, Kashiwab-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.
15. Braun RP, Gutkowicz-Kusins D, Rabinirovitz 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.
16. Gaud S, Zarandona JM, Raab SS, English K, Tardik JM. 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.
17. Patrawala S, Maley A, Greskovic 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.
18. Niebelng MG, Haydu LE, Karim ZR, Thompson JF, Scoller 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 2016;23:2245-51. doi:10.1245/s10434-016-3862-x.
19. Komor ER, Goinin 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.
20. Barnhill RL, Surber C, From L, et al. Atypical Spitz nevi/tumors: loss of consensus for diagnosis, discrimination from melanoma, and prediction of outcome. Hum Pathol 1999;30:513-20. doi:10.1016/S0046-8177(99)00193-4.
21. Meyer LJ, Piepko 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)00061-3.
22. Ferra J, 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.
23. Collody-Franchi A, Frederic J, et al. 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.1711630107.
24. Miller LC, 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.d0.100264.x.
25. Lock Andersen J, Hou-Jensen K, Hansen JP, Jensen NK, Søgaard A, Andersen PK. Observer variation in histological classification of cutaneous malignant melanoma. Scand J Plast Reconstr Surg Hand Surg 1995;29:141-8. doi:10.1080/08941995.1995.1160648.
26. Spatz A, Cook MG, Elder DE, Piepko 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 Sculier 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, Groschans 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 multivariable, population-based analysis of histologic dysplasia in melanocytic nevi. *J Am Acad Dermatol* 1994;30:707-14. doi:10.1016/S0190-9622(94)70054-5.

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.19970389040019002.

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, Joo 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. *Pathology 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:10.1111/j.1365-2559.1996.tb09718.x.

54 McCullough BJ, Johnson GR, Martin BJ, Janik 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