Published in final edited form as:

Histopathology. 2014 August ; 65(2): 240–251. doi:10.1111/his.12387.

Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel

Kimberly H Allison, Lisa M Reisch¹, Patricia A Carney², Donald L Weaver³, Stuart J Schnitt⁴, Frances P O'Malley⁵, Berta M Geller⁶, and Joann G Elmore¹

¹Department of Pathology, University of Washington Medical Center, Seattle, WA, USA, Department of Internal Medicine, University of Washington School of Medicine, Seattle, WA, USA

²Departments of Family Medicine and Public Health & Preventive Medicine, Oregon Health & Science University, Portland, OR, USA

³Department of Pathology, University of Vermont College of Medicine and Vermont Cancer Center, Burlington, VT, USA

⁴Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA

⁵Department of Laboratory Medicine, Keenan Research Centre of the Li Ka Shing Knowledge Institute, St Michael's Hospital, University of Toronto, Toronto, ON, Canada

⁶Department of Family Medicine, Health Promotion Research, University of Vermont, Burlington, VT, USA

Abstract

Aims—To gain a better understanding of the reasons for diagnostic variability, with the aim of reducing the phenomenon.

Methods and results—In preparation for a study on the interpretation of breast specimens (B-PATH), a panel of three experienced breast pathologists reviewed 336 cases to develop consensus reference diagnoses. After independent assessment, cases coded as diagnostically discordant were discussed at consensus meetings. By the use of qualitative data analysis techniques, transcripts of 16 h of consensus meetings for a subset of 201 cases were analysed. Diagnostic variability could be attributed to three overall root causes: (i) pathologist-related; (ii) diagnostic coding/study methodology-related; and (iii) specimen-related. Most pathologist-related root causes were attributable to professional differences in pathologists' opinions about whether the diagnostic criteria for a specific diagnosis were met, most frequently in cases of atypia. Diagnostic coding/study methodology-related root causes were primarily miscategorizations of descriptive text diagnoses, which led to the development of a standardized electronic diagnostic form (BPATH-

Address for correspondence: K H Allison, MD, Department of Pathology, Stanford University Medical Center, 300 Pasteur Drive, Lane L235, Stanford, CA 94305, USA. allisonk@stanford.edu.

Author contributions

All authors listed have contributed to this work. Kimberly Allison and Lisa Reich performed the qualitative review/data collection, data analysis, and manuscript preparation. Patricia Carney and Berta Gellar assisted with data collection and manuscript editing. Stuart Schnitt, Frances O'Malley and Donald Weaver were the pathologists involved in the study, and also assisted with manuscript editing. Joann Elmore and Donald Weaver assisted with the review and oversight of data collection, data analysis, and manuscript editing.

Dx). Specimen-related root causes included artefacts, limited diagnostic material, and poor slide quality. After re-review and discussion, a consensus diagnosis could be assigned in all cases.

Conclusions—Diagnostic variability is related to multiple factors, but consensus conferences, standardized electronic reporting formats and comments on suboptimal specimen quality can be used to reduce diagnostic variability.

Keywords

atypical ductal hyperplasia; borderline breast lesions; breast pathology; diagnostic disagreement; diagnostic variability; pathologist agreement

Introduction

In the clinical practice of medicine, differences in opinion on both diagnosis and treatment are generally accepted as part of the professional art, despite efforts to standardize care. In surgical pathology, tissue-based diagnoses are considered to constitute the diagnostic gold standard, and are used to categorize, treat and study disease; differences in diagnostic opinion and diagnostic variation are often perceived as unacceptable. Unfortunately, the reasons underlying diagnostic variation in pathology are poorly understood, and discussing differences in diagnostic opinion frequently raises the spectre of medical liability.

Although diagnostic agreement is high in many critical areas of pathology, high levels of variability have been identified in diagnoses on the borderlines between established entities.^{1–15} Studies focusing specifically on breast pathology have indicated that diagnostic agreement varies with diagnostic category, with high diagnostic agreement for benign disease, high-grade ductal carcinoma *in situ* (DCIS), and invasive carcinoma, but lower agreement for borderline categories such as atypical ductal hyperplasia (ADH) and limited-extent low-grade DCIS.^{16–28} ADH is frequently treated with excisional biopsy alone, whereas DCIS is treated with complete surgical removal, and often with additional radiation and hormonal therapy. Because major clinical treatment thresholds exist between a diagnosis of ADH and a diagnosis of DCIS, the high diagnostic variability for these two entities has been highlighted by many studies, and has been a focus of negative media attention suggesting that pathologists are ‘prone to error’.^{18,21,25,26,29–31}

Both pathologists who specialize in breast pathology and general surgical pathologists have high levels of diagnostic variability in these challenging areas.²³ Some studies have suggested ways to reduce interobserver variability in breast pathology; however, strong agreement on ADH versus limited-extent low-grade DCIS remains elusive.^{15,21,26,27,29,32} Although many authors have speculated that better diagnostic criteria and improved training in breast pathology are needed to improve agreement, few studies have focused on the root causes of diagnostic variability.^{15,30,33} As experienced breast pathologists also have high levels of diagnostic variability in this area, the factors underlying variability cannot be related only to level of experience, and warrant further explanation.

As part of the Breast Pathology (B-PATH) study, an NIH-funded study examining diagnostic variability in breast pathology, a panel of three experienced breast pathologists

(F.O.M., S.J.S., and D.L.W.) was convened to establish consensus reference diagnoses on a series of breast pathology cases.³⁴ All cases were reviewed and interpreted independently by each pathologist, and any cases coded as initial diagnostic discordances were reviewed at a series of consensus conferences that were recorded by the facilitator. During consensus conference discussions, it became clear that cases coded as discordant had a range of underlying reasons for diagnostic variability, some of which the experienced pathologists did not feel were true ‘disagreements’. This prompted a formal qualitative content analysis of the transcripts from these consensus meetings, with the goal of identifying and describing underlying themes for diagnostic variability among specialist breast pathologists. Although this study is limited to the issues raised by these particular breast pathologists, the themes and reasons for diagnostic variability illuminated by the study illustrate the complexities underlying diagnostic variability.

Methods

The larger B-PATH study and test set development have been described elsewhere.³⁴ In brief, 336 cases were randomly selected for inclusion in the study, but were enriched for statistical power to include a higher proportion of ADH and DCIS cases than might be typically encountered in clinical practice. A consensus diagnosis on each case was established through a three-step process: (i) each of the three experienced pathologists reviewed all cases independently on the same set of glass slides, blinded to each other’s interpretations, and entered their independent diagnoses into a diagnostic assessment form (see below for details of the diagnostic form development); (ii) each independent diagnosis was coded into one of five hierarchical diagnostic assessment categories [invasive, DCIS, atypical (ADH and intraductal papilloma with atypia), proliferative without atypia, and non-proliferative]; and (iii) all cases with categorical diagnostic discordance were reviewed at a multi-headed microscope and discussed in person at a series of four consensus conferences, with the goal of reaching diagnostic agreement on all cases. Among the initial 336 cases, the final consensus diagnosis breakdown was as follows: invasive, $N = 43$ (12.8%); DCIS, $N = 88$ (26.2%); atypia, $N = 72$ (21.4%); proliferative without atypia, $N = 96$ (28.6%); and non-proliferative, $N = 37$ (11%).

DEVELOPING THE BPATH-DX DIAGNOSTIC ASSESSMENT FORM

The final diagnostic assessment form (BPATH-Dx; Breast Pathology Assessment Tool and Hierarchy for Diagnosis) developed for the study underwent several revisions. Before the first and second consensus meetings, independent diagnostic assessments were performed with paper forms that included a list of standard diagnoses in a check-box format and an ‘other’ diagnosis category with a field for handwritten descriptive diagnoses. On the basis of what was learned at the first two consensus meetings, modifications were made to the organization and list of diagnostic categories on the forms to specifically capture diagnoses that were initially coded as ‘other’. In addition, a new section was added to indicate whether a case was considered to be ‘borderline’ between two diagnoses. Finally, the paper format was replaced with an electronic form with direct data entry. The electronic form required that each case was allocated to a diagnostic assessment category, with no ‘other’ diagnosis check-box, but it did include an option for additional free-text comments on a case. The final

BPATH-Dx form (Figure 1) was finalized prior to the third expert consensus meeting, and is being used in the main B-PATH study of diagnostic variation.

CONSENSUS MEETINGS

The consensus meetings included an experienced facilitator (P.C. or B.G.), and used a modified Delphi approach to guide the process of reaching agreement.³⁵ This involved compiling the independent reviews for each case, providing the three experts with their initial interpretations, and a re-review of the slides at a multi-headed microscope followed by a facilitated discussion. The facilitated discussion continued until a final consensus reference diagnosis was reached for the case among all three expert pathologists. The consensus meetings were audio-recorded and later transcribed for qualitative review. The University of Washington's Institutional Review Board approved the study, and all three breast pathology experts consented to participate.

CONTENT ANALYSIS OF TRANSCRIPTS

Transcripts from 16 h of consensus meetings were independently analysed using classic content analysis and coded by two authors (K.A. and L.R.) utilizing the following steps: (i) search for emerging themes; (ii) independent coding; (iii) consensus coding; and (iv) quantifying codes within each theme.^{36–39} By the use of these steps, an iterative process was established to reach joint agreement on key themes, definitions, and root causes of disagreement (Table 1). Key themes considered to be the root causes of the disagreement for each case discussed were tallied, and quotations were identified that exemplified the themes. Transcripts of the first consensus meeting were selected, because it represented the first data form used. Consensus meeting transcripts were then evaluated from the third meeting, because modifications to the BPATH-Dx form had been finalized. After reviewing the transcripts of these two full-day consensus meetings (the first and the third), the investigators agreed that thematic saturation had been reached; additional transcript review was not performed.

Results

The transcripts reviewed for this qualitative study to identify themes for diagnostic variability included 16 h of consensus discussion after independent expert review of 201 (59.8%) of the 336 total cases independently assessed by the three expert breast pathologists for the B-PATH study. The consensus final diagnosis categories for these 201 cases reviewed included: invasive ($N = 31$; 15.4%), DCIS ($N = 41$; 20.4%), atypia ($N = 34$; 16.9%), proliferative without atypia ($N = 64$; 31.8%), and non-proliferative ($N = 30$; 15.9%), with one case (0.5%) not applicable (excluded before a consensus diagnosis was reached). Of the 201 source cases, 75 were initially coded as categorically discordant and formed the focus of the consensus discussions analysed.

THEMES IDENTIFIED AND ROOT CAUSES OF DIAGNOSTIC VARIABILITY

We identified 14 key themes of diagnostic variability; these could be grouped into one of three overall root causes of diagnostic variability: (i) pathologist-related; (ii) diagnostic coding/study methodology-related; and (iii) specimen-related (Table 1). The most common

root cause of diagnostic variability was diagnostic coding/study methodology-related, followed by pathologist-related and specimen-related causes. The majority of atypia diagnoses, which included both ADH and intraductal papilloma with atypia, were associated with pathologist-related themes for variability, the most common of which comprised subtle differences in professional opinion about whether the features present met the criteria for a specific diagnosis. In contrast, for all remaining diagnostic categories, initial diagnostic variability was most commonly caused by diagnostic coding/study methodology-related themes. The frequency rankings of the key themes within each root cause category are shown in Table 1 and are further described, with illustrative quotations, in Table 2.

PATHOLOGIST - RELATED REASONS FOR DIAGNOSTIC VARIABILITY

Of the pathologist-related reasons for diagnostic variability, the most common underlying theme comprised subtle differences in professional opinion regarding whether the features present met the diagnostic criteria for a specific diagnosis. Discussion at consensus meetings about these cases focused on the concept of predictable diagnostic variability, because of the borderline nature of the histological changes present. These discussions did not involve questions about invasive cancer or high-grade DCIS, but most frequently involved cases with a question regarding the presence of atypia versus either usual ductal hyperplasia (UDH) or limited-extent low-grade DCIS. The pathologists all typically agreed on the diagnostic criteria, morphological features present and differential diagnosis of these cases. However, they had subtle differences of professional opinion on whether the specific morphological features present met the diagnostic criteria; these often related to an individual pathologist's threshold for a particular diagnosis in a specific case. For example, discussion about the extent of the atypical morphological features present and how well developed they were architecturally often revealed subtle differences in diagnostic thresholds (see Figure 2 for examples).

The second most common pathologist-related reason for disagreement was when one or two panelists did not note a focal finding noted by another panelist during their independent review of the case. During discussion of these cases, a pathologist verbally acknowledged scanning over a focal diagnostic finding, and subsequently agreed with another pathologist's diagnosis. In all of these cases, the finding of interest was noted to be only focal (two were focal areas of atypical lobular hyperplasia, two were focal ADH, one was focal ADH in a papilloma, and one was focal microinvasion in a background of DCIS). In consensus review, reference to the diagnostic coding/study methodology-related setting of the slide review was discussed as a possible influence when focal findings were missed. In addition, in contrast to normal clinical settings, diagnoses were based on review of a single slide rather than multiple levels or additional slides, which may have provided additional opportunity to observe a focal finding.

In addition, the use of different diagnostic philosophies (making a diagnosis on the basis of morphological criteria alone versus making a diagnosis on the basis of both morphology and clinical context) was the third most frequent cause for diagnostic variability for the pathologist-related cases. Discussion of these cases centred on differences in the utilization of available clinical data (e.g. patient age and type of biopsy versus other specimen type)

and the potential clinical impact of a diagnosis as opposed to basing a diagnosis purely on morphological features. These cases typically involved differences in the clinical impact of a core biopsy versus excisional biopsy diagnosis, and differences in the way in which institutions manage diagnoses of high-risk lesions on core needle biopsy.

Using different diagnostic criteria or noting different diagnostic features both ranked fourth among the causes of pathologist-related diagnostic variability. Differing criteria were primarily related to diagnoses of flat epithelial atypia (FEA). Finally, acknowledging fatigue as a reason for variability was a cause of pathologist-related variability, but was the least frequently identified pathologist-related root cause.

DIAGNOSTIC CODING/STUDY METHODOLOGY-RELATED REASONS FOR DIAGNOSTIC VARIABILITY

Diagnostic coding/study methodology-related reasons for diagnostic variability were considered during consensus discussion as a source of ‘false diagnostic discordance’. These reasons were primarily related to the diagnostic assessment forms, coding, and diagnostic hierarchies used in the study. Many of these issues were addressed during the course of the study to reduce diagnostic discordance.

The cases discussed at the first consensus meeting were individually assessed with an initial version of the BPATH-Dx assessment form that allowed for handwritten diagnoses to be listed in an ‘other’ diagnosis section; a third of these earlier cases were noted to be discordant primarily because of miscoding; it was clear during consensus review that these cases did not represent true diagnostic disagreements (see Figure 3 for an example). The ‘other’ diagnosis category was subsequently eliminated from the diagnostic assessment form, and additional diagnostic categories were added. In addition, the form was changed from a paper to an electronic format that required selection of a listed diagnostic category (rather than the pathologist being able to write in an unlisted diagnosis).

The electronic form without an option for ‘other’ diagnoses greatly reduced discordance related to miscoding of handwritten diagnoses; however, this change produced an increase in diagnostic variability related to the pathologist feeling forced to choose an available diagnostic category. In some instances, pathologists mentioned choosing their final diagnostic category out of frustration, because there was no option for a diagnosis with wording that they preferred to use in clinical practice. This was typical for cases with features that were borderline between two diagnoses, where the pathologists wanted their diagnoses to better reflect their equivocation. Because of this finding, the option of flagging a case as ‘borderline’ and noting the differential diagnoses was added to the electronic form check-boxes.

As detailed in the Methods section, initial discordances were identified by use of a diagnostic hierarchy that grouped diagnoses with similar risk. For example, ADH and ‘intraductal papilloma with atypia’ were in the same hierarchical category, so these two diagnoses would not have been flagged as a diagnostic discordance (see Figure 1 footnote). The hierarchical categories were modified slightly as the study progressed. A consequence of this decision was that earlier cases initially coded and discussed during the consensus

meeting as discordant were not coded as discordant later in the study. The majority of these cases involved a diagnosis of FEA as the most severe diagnostic finding. FEA was initially coded in the atypia category, but because FEA was not a diagnosis in our random selection strategy for the study sets, and because of its low risk for future carcinoma, it was later placed in the proliferative without atypia category. There are plans to analyse the data in the future in both ways for cases with FEA as the highest-order diagnosis. However, it was notable that how a specific lesion is coded as belonging to a diagnostic category can be a diagnostic coding/study methodology-related reason for perceived diagnostic variability.

Some diagnostic discordance was related to the proliferative without atypia category versus the non-proliferative category, a difference that was discussed as being clinically irrelevant in most cases. A common cause of diagnostic variability in this category was related to the presence or absence of columnar cell hyperplasia. These discussions were helpful in planning future analyses, as we decided to merge non-proliferative and proliferative diagnoses together into one clinically similar category for some statistical analyses.

SPECIMEN - RELATED REASONS FOR DIAGNOSTIC VARIABILITY

Poor slide quality or artefacts accounted for the majority of specimen-related diagnostic variability, including five cases that were eventually excluded from the main study because their poor quality severely compromised the ability to make a diagnosis based on the single slide. These artefacts included features such as severe chatter, holes in the tissue, severe cautery, and poor stain quality.

A less common specimen-related reason for variability was limited diagnostic material present on the slide.

Discussion

Diagnostic variability in pathology occurs, but the underlying reasons are poorly understood. We identified many potential causes of perceived diagnostic disagreement during our formal qualitative review of the process of establishing consensus reference diagnoses of breast pathology cases. Three main types of root cause were identified: (i) pathologist-related; (ii) diagnostic coding/study methodology-related; and (iii) specimen-related. Of these, pathologist-related root causes were the most common.

Our analysis suggests that many pathologist-related disagreements may represent valid differences in professional opinion, and should not be perceived as medical errors. Features considered to be borderline between two diagnoses were predictable causes of variability. Subtle differences in professional opinion can become critical in these borderline cases. This was particularly common with cases of atypia (ADH and papilloma with atypia). Our experienced pathologists typically used the same diagnostic criteria and noted the same morphological features in a case, but sometimes had professional differences of opinion about whether or not the features present met a diagnostic threshold. Possible explanations for this phenomenon include differences in pathologist training and regional clinical practice patterns.

It should be acknowledged that borderline categories such as ADH and low-grade DCIS may be prone to diagnostic variability because of their innate lack of distinction from a biological perspective, despite their clinical utility as diagnostic categories. For example, ADH harbours similar molecular aberrations to those found in low-grade DCIS, indicating a nearly identical biological genesis.⁴⁰ This morphological and genetic biological similarity has also been observed in other borderline diagnostic categories of pathology, such as atypical versus non-atypical complex endometrial hyperplasia.⁴¹

Pathology is not the only field in medicine with variability around borderline areas and thresholds.⁴² For example, studies have indicated physician diagnostic disagreement rates as high as 33% in the diagnosis of acute coronary syndrome.⁴³ Diagnostic and treatment variability for borderline cases in medicine is to be expected, and should be acknowledged. In our study, the experienced breast pathologists found it very useful to be able to categorize a case as borderline and provide a differential diagnosis. We will have the same feature available to all participants in the larger study. If the existence of borderline cases in clinical practice is acknowledged, clinical decisions may be made with more appropriate thought and context rather than with the misconception that the case is a straightforward example of a particular diagnosis or that the diagnosis of one pathologist is 'right' and that of another is 'wrong'. This more pragmatic approach to how a particular lesion should be managed rather than assuming that a diagnosis fits perfectly into a morphology-based diagnostic category could be the basis for developing more practical management-related diagnostic categories.

We may have overestimated the extent of pathologist-related diagnostic variability, given the challenging case mix and lack of access to clinical context, imaging correlation, additional levels, adjunctive immunostains and second opinions that may decrease variability. On the other hand, the pathologists were interpreting only a single slide, which might result in higher levels of agreement. Other studies have also suggested that sets of training slides or images and review of specific diagnostic criteria can help to reduce diagnostic discordance in breast lesions in the study setting. In an attempt to mimic a typical practice setting, our study did not include training sets or review of specific diagnostic criteria. However, diagnostic agreement was reached in all cases after consensus review discussion, supporting consensus conferences as another useful method for decreasing diagnostic discordance in clinical practice. Although, in the current expert review study, only a single set of glass slides was used, the larger B-PATH study will also examine diagnostic discordance using digital versus glass slides, as digital slides may also constitute a useful methodology for rapid consultation in borderline cases. Studies of the impact of second reviews, consensus discussions, digital pathology consultations and reporting of borderline features should be encouraged, to help establish standards that can improve diagnostic and management consensus in clinical practice.

Diagnostic coding/study methodology-related reasons for diagnostic discordance were also noted in our study, and resulted in modification of our study design and the development of the BPATH-Dx diagnostic forms. The use of our BPATH-Dx standardized form and automated electronic data entry led to improved diagnostic agreement. These findings regarding the importance of standardized forms have implications for how studies on diagnostic agreement are designed, implemented, and reported, so that disagreement rates

are not artificially inflated and clinically relevant subtleties are better understood. This also has clinical relevance, because descriptive diagnosis reporting in clinical practice may lead to misinterpretation by treating clinicians. The counter-argument is that standardized electronic diagnostic forms force pathologists into choosing specific diagnostic categories when more descriptive diagnoses or alternative terminology may be more appropriate in guiding therapy. This issue can be partially addressed by having a reporting mechanism that allows the annotation of a case as borderline between two diagnoses.

Research studies, coding for billing and tumour registries rely on accurate diagnostic coding, so standardized diagnostic taxonomies with clear definitions and automated electronic data entry programs are often employed to reduce errors in diagnostic categorization. Clinical surgical pathology reports may also benefit from standardization for clarity in translation to referring providers and patients, for coding and billing, and for research. The College of American Pathologists (CAP) supplies checklists and templates for invasive breast cancer diagnoses that have increased the use of standardized reporting.⁴⁴ However, specimens such as breast needle core biopsies that include risk lesions such as ADH do not have clear reporting standards in North America. This stands in contrast to our European colleagues, whose guidelines for quality assurance recommend the use of a pathology categorization system (B1–B5) for reporting breast biopsy diagnoses.^{45,46} This system is somewhat analogous to radiologists' use of the Breast Imaging Reporting and Diagnosis System (BI-RADS), which classifies the radiologists' diagnostic findings into five risk categories. The development of a similar schema in the USA for breast biopsies presents some challenges, which we also faced in our study, but could help to reduce diagnostic variability. Our BPATH-Dx assessment tool is currently being used in clinical practice in a northwestern medical health system (H. Nelson, verbal communication), where a natural language-processing tool organizes pathology reports into a database that includes categorization of all diagnoses, including non-malignant diagnoses.

Not surprisingly, we also found that poor specimen quality or limited tissue quantity leads to diagnostic variability. Artefacts and severely limited diagnostic material were common underlying sources of diagnostic variability, and, after consensus conference review, these cases were excluded from the study if the material was deemed to be inadequate for diagnosis by the expert pathologists. Limited diagnostic material has been similarly observed to affect diagnostic agreement in other areas of pathology.⁴¹ Limited diagnostic material is a problem that will vary by laboratory and can affect diagnostic agreement. Suboptimal specimen characteristics should be documented in clinical reports when the accuracy of a diagnosis may be affected.

Our study was limited to three experienced pathologists and a single clinical topic. The themes that emerged from our qualitative review may be dependent on the cases included in the study, which were selected to include more cases of atypia than might typically be encountered in daily practice. In addition, the study pathologists were reviewing a single slide, and did not have a clinical context other than patient age. However, our sampling strategy allowed us to include cases common to routine practice, rather than specifically selecting cases because of their level of difficulty or exemplary histology. A major strength of this analysis is the rigorous review of the complexities of diagnostic agreement among

breast pathology experts for a set of cases randomly drawn from community practice. A quantitative analysis of diagnostic agreement for the experienced breast pathologists, as well as for the larger BPATH study (which includes predominantly community-based pathologists), will also be performed to further examine issues in diagnostic variability for specific breast pathology diagnostic categories.

In summary, our observations provide an improved understanding of the challenges involved in achieving and accurately and appropriately reporting levels of diagnostic agreement in pathology. Our findings provide useful insights for improving quality assurance in both research and clinical settings. This study could serve as a framework for the evaluation of observer variability in future studies addressing this issue in both breast pathology and other areas of pathology.

Acknowledgments

The National Cancer Institute (R01 CA140560, KO5 CA104699) supported this work. The authors wish to thank Tom Morgan for his assistance with data analysis.

References

1. Abdollahi A, Meysamie A, Sheikhabaei S, et al. Inter/intraobserver reproducibility of Gleason scoring in prostate adenocarcinoma in Iranian pathologists. *Urol J*. 2012; 9:486–490. [PubMed: 22641492]
2. Barnhill RL, Argenyi Z, Berwick M, et al. Atypical cellular blue nevi (cellular blue nevi with atypical features): lack of consensus for diagnosis and distinction from cellular blue nevi and malignant melanoma ('malignant blue nevus'). *Am J Surg Pathol*. 2008; 32:36–44. [PubMed: 18162768]
3. Bean SM, Meara RS, Vollmer RT, et al. p16 Improves interobserver agreement in diagnosis of anal intraepithelial neoplasia. *J Low Genit Tract Dis*. 2009; 13:145–153. [PubMed: 19550211]
4. Coco DP, Goldblum JR, Hornick JL, et al. Interobserver variability in the diagnosis of crypt dysplasia in Barrett esophagus. *Am J Surg Pathol*. 2011; 35:45–54. [PubMed: 21164286]
5. El-Zimaity HM, Wotherspoon A, de Jong D. Interobserver variation in the histopathological assessment of MALT/MALT lymphoma: towards a consensus. *Blood Cells Mol Dis*. 2005; 34:6–16. [PubMed: 15607694]
6. Fadare O, Parkash V, Dupont WD, et al. The diagnosis of endometrial carcinomas with clear cells by gynecologic pathologists: an assessment of interobserver variability and associated morphologic features. *Am J Surg Pathol*. 2012; 36:1107–1118. [PubMed: 22790851]
7. Fleskens SA, Bergshoeff VE, Voogd AC, et al. Interobserver variability of laryngeal mucosal premalignant lesions: a histopathological evaluation. *Mod Pathol*. 2011; 24:892–898. [PubMed: 21499237]
8. Foss FA, Milkins S, McGregor AH. Inter-observer variability in the histological assessment of colorectal polyps detected through the NHS Bowel Cancer Screening Programme. *Histopathology*. 2012; 61:47–52. [PubMed: 22486166]
9. Gerhard R, da Cunha Santos G. Inter- and intraobserver reproducibility of thyroid fine needle aspiration cytology: an analysis of discrepant cases. *Cytopathology*. 2007; 18:105–111. [PubMed: 17397495]
10. Gupta T, Nair V, Epari S, Pietsch T, Jalali R. Concordance between local, institutional, and central pathology review in glioblastoma: implications for research and practice: a pilot study. *Neurol India*. 2012; 60:61–65. [PubMed: 22406782]
11. Kerkhof M, van Dekken H, Steyerberg EW, et al. Grading of dysplasia in Barrett's oesophagus: substantial interobserver variation between general and gastrointestinal pathologists. *Histopathology*. 2007; 50:920–927. [PubMed: 17543082]

12. Longnecker DS, Adsay NV, Fernandez-del Castillo C, et al. Histopathological diagnosis of pancreatic intraepithelial neoplasia and intraductal papillary-mucinous neoplasms: interobserver agreement. *Pancreas*. 2005; 31:344–349. [PubMed: 16258368]
13. Parkash V, Bifulco C, Feinn R, Concato J, Jain D. To count and how to count, that is the question: interobserver and intraobserver variability among pathologists in lymph node counting. *Am J Clin Pathol*. 2010; 134:42–49. [PubMed: 20551265]
14. Puppa G, Senore C, Sheahan K, et al. Diagnostic reproducibility of tumour budding in colorectal cancer: a multicentre, multinational study using virtual microscopy. *Histopathology*. 2012; 61:562–575. [PubMed: 22765314]
15. Stang A, Trocchi P, Ruschke K, et al. Factors influencing the agreement on histopathological assessments of breast biopsies among pathologists. *Histopathology*. 2011; 59:939–949. [PubMed: 22092405]
16. Anderson TJ, Sufi F, Ellis IO, Sloane JP, Moss S. Implications of pathologist concordance for breast cancer assessments in mammography screening from age 40 years. *Hum Pathol*. 2002; 33:365–371. [PubMed: 11979379]
17. Beck JS. Observer variability in reporting of breast lesions. *J Clin Pathol*. 1985; 38:1358–1365. [PubMed: 3001153]
18. Bianchi S, Palli D, Galli M, et al. Reproducibility of histological diagnoses and diagnostic accuracy of non palpable breast lesions. *Pathol Res Pract*. 1994; 190:69–76. [PubMed: 8065991]
19. Bodian CA, Perzin KH, Lattes R, Hoffmann P. Reproducibility and validity of pathologic classifications of benign breast disease and implications for clinical applications. *Cancer*. 1993; 71:3908–3913. [PubMed: 8508356]
20. Collins LC, Connolly JL, Page DL, et al. Diagnostic agreement in the evaluation of image-guided breast core needle biopsies: results from a randomized clinical trial. *Am J Surg Pathol*. 2004; 28:126–131. [PubMed: 14707874]
21. Elston CW, Sloane JP, Amendoeira I, et al. Causes of inconsistency in diagnosing and classifying intraductal proliferations of the breast. European Commission Working Group on Breast Screening Pathology. *Eur J Cancer*. 2000; 36:1769–1772. [PubMed: 10974624]
22. Ghofrani M, Tapia B, Tavassoli FA. Discrepancies in the diagnosis of intraductal proliferative lesions of the breast and its management implications: results of a multinational survey. *Virchows Arch*. 2006; 449:609–616. [PubMed: 17058097]
23. Palazzo JP, Hyslop T. Hyperplastic ductal and lobular lesions and carcinomas in situ of the breast: reproducibility of current diagnostic criteria among community- and academic-based pathologists. *Breast J*. 1998; 4:230–237. [PubMed: 21223441]
24. Palli D, Galli M, Bianchi S, et al. Reproducibility of histological diagnosis of breast lesions: results of a panel in Italy. *Eur J Cancer*. 1996; 32A:603–607. [PubMed: 8695260]
25. Rosai J. Borderline epithelial lesions of the breast. *Am J Surg Pathol*. 1991; 15:209–221. [PubMed: 1847606]
26. Schnitt SJ, Connolly JL, Tavassoli FA, et al. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. *Am J Surg Pathol*. 1992; 16:1133–1143. [PubMed: 1463092]
27. Sloane JP, Ellman R, Anderson TJ, et al. Consistency of histopathological reporting of breast lesions detected by screening: findings of the UK National External Quality Assessment (EQA) Scheme. UK National Coordinating Group for Breast Screening Pathology. *Eur J Cancer*. 1994; 30A:1414–1419. [PubMed: 7833094]
28. Wells WA, Carney PA, Eliassen MS, Tosteson AN, Greenberg ER. Statewide study of diagnostic agreement in breast pathology. *J Natl Cancer Inst*. 1998; 90:142–145. [PubMed: 9450574]
29. Jain RK, Mehta R, Dimitrov R, et al. Atypical ductal hyperplasia: interobserver and intraobserver variability. *Mod Pathol*. 2011; 24:917–923. [PubMed: 21532546]
30. Perkins C, Balma D, Garcia R. Why current breast pathology practices must be evaluated. A Susan G. Komen for the Cure white paper: June 2006. *Breast J*. 2007; 13:443–447. [PubMed: 17760664]
31. Saul S. Prone to error: earliest steps to find cancer. *New York Times*. 2010:1.
32. Sloane JP, Amendoeira I, Apostolikas N, et al. Consistency achieved by 23 European pathologists in categorizing ductal carcinoma in situ of the breast using five classifications. *European*

- Commission Working Group on Breast Screening Pathology. *Hum Pathol.* 1998; 29:1056–1062. [PubMed: 9781641]
33. Masood S, Rosa M. Borderline breast lesions: diagnostic challenges and clinical implications. *Adv Anat Pathol.* 2011; 18:190–198. [PubMed: 21490436]
34. Oster NV, Carney PA, Allison KH, et al. Development of a diagnostic test set to assess agreement in breast pathology: practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *BMC Women's Health.* 2013; 13:3. [PubMed: 23379630]
35. Hsu CC, Sanford BA. The Delphi technique: making sense of consensus. *Pract Assess Res Eval.* 2007; 12:1–8.
36. Borkan, J. Immersion/crystallization. In: Crabtree, BF.; Miller, WL., editors. *Doing qualitative research.* Thousand Oaks, CA: Sage Publications; 1999. p. 179-194.
37. Miller W, Crabtree BF. Qualitative analysis: how to begin making sense. *Fam Pract Res J.* 1994; 14:289–297. [PubMed: 7976480]
38. Miller, W.; Crabtree, B.; Denzin, N.; Lincoln, Y. *Handbook of qualitative research.* Newbury Park, CA: Sage Publications; 1994.
39. Bauer, MW. Classical content analysis: a review. In: Bauer, MW.; Gaskell, G., editors. *Qualitative researching with text, image and sound.* New York, NY: Sage Publishing; 2000. p. 131-151.
40. Lakhani SR, Collins N, Stratton MR, Sloane JP. Atypical ductal hyperplasia of the breast: clonal proliferation with loss of heterozygosity on chromosomes 16q and 17p. *J Clin Pathol.* 1995; 48:611–615. [PubMed: 7560165]
41. Allison KH, Reed SD, Voight LF, Jordan CD, Newton KM, Garcia RL. Diagnosing endometrial hyperplasia: why is it so difficult to agree? *Am J Surg Pathol.* 2008; 32:691–698. [PubMed: 18347507]
42. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (Final Installment). *J Clin Epidemiol.* 1992; 45:567–580. [PubMed: 1607896]
43. Nagurney JT, Brown DF, Chae C, et al. Disagreement between formal and medical record criteria for diagnosis of acute coronary syndrome. *Acad Emerg Med.* 2005; 12:446–452. [PubMed: 15860697]
44. Lester SC, Bose S, Chen YY, et al. Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Arch Pathol Lab Med.* 2009; 133:1515–1538. [PubMed: 19792042]
45. Pagni F, Bosisio FM, Salvioni D, Colombo P, Leone BE, Di Bella C. Application of the British National Health Service Breast Cancer Screening Programme classification in 226 breast core needle biopsies: correlation with resected specimens. *Ann Diagn Pathol.* 2012; 16:112–118. [PubMed: 22056037]
46. Ellis IO, Humphreys S, Michell M, Pinder SE, Wells CA, Zakhour HD. Best Practice No. 179. Guidelines for breast needle core biopsy handling and reporting in breast screening assessment. *J Clin Pathol.* 2004; 57:897–902. [PubMed: 15333647]

BPATH-Dx Diagnostic Assessment Form

I. Histologic Assessment: Diagnoses—Check all that apply. Choose the best fit among the options.

Non-Proliferative changes

- ☐ Non-Proliferative changes only

Proliferative lesion without atypia

- ☐ Fibroadenoma
☐ Intraductal papilloma without atypia
☐ Usual ductal hyperplasia
☐ Columnar cell hyperplasia/ Columnar cell change
☐ Sclerosing adenosis
☐ Radial scar/ complex sclerosing lesion

Atypical lesion

- ☐ Flat epithelial atypia
☐ Atypical ductal hyperplasia
☐ Intraductal papilloma with atypia
☐ Atypical lobular hyperplasia

Carcinoma in situ

- ☐ Ductal carcinoma in situ:

Nuclear grade ☐ Low ☐ Intermediate ☐ High

Necrosis ☐ Absent ☐ Present, focal (small foci/ single cell necrosis) ☐ Present, central (expansive "comedo" necrosis)

- ☐ Lobular carcinoma in situ (for mixed ductal & lobular features, check both DCIS & LCIS boxes and nuclear grade + necrosis)

Invasive carcinoma

- ☐ Invasive carcinoma (ductal, lobular or other special type):

a. Tubule formation score ☐ 1 ☐ 2 ☐ 3

b. Nuclear grade score ☐ 1 ☐ 2 ☐ 3

c. Mitotic activity score ☐ 1 ☐ 2 ☐ 3

Overall Nottingham grade ☐ Low (total score 3-5) ☐ Intermediate (6,7) ☐ High (8,9)

Additional comments (optional) _____

II. Do you consider your most advanced diagnosis for this case to be borderline between two diagnoses?

☐ Yes* ☐ No

* If yes, responders were prompted to provide the two diagnoses that the case was borderline between using the same options as above

What particular features made you favor the final diagnostic category you chose for the lesion? (Optional)

FOOTNOTE: Final Diagnostic Hierarchical categories

Coded Diagnosis	Maps To The Following Diagnostic Category
Invasive (ductal, lobular, other special type)	Invasive
Ductal carcinoma in situ	DCIS
Atypical ductal hyperplasia	Atypical
Intraductal papilloma with atypia	Atypical
Usual ductal hyperplasia	Proliferative
Columnar cell change/hyperplasia	Proliferative
Sclerosing Adenosis	Proliferative
Radial Scar / complex Sclerosing lesion	Proliferative
Flat epithelial atypia	Proliferative
Intraductal papilloma w/o atypia	Proliferative
Non-Proliferative only	Non-Proliferative
Fibroadenoma	Non-Proliferative
LCIS*	Please see footnotes
ALH*	Please see footnotes

*Because cases were not selected for a diagnosis of lobular in situ neoplasia, if ALH or LCIS were diagnosed, the case mapped to the other diagnoses also on the slide or is classified as non-proliferative if no other diagnoses are noted.

Figure 1.

Final electronic BPATH-Dx diagnostic form and diagnostic hierarchical categories. Paper diagnostic forms were replaced with the final BPATH-Dx standardized electronic data entry form. During evaluation for pathologist agreement, the highest-order diagnosis present in an individually scored BATH-Dx form was coded according to the diagnostic hierarchical categories shown in the footnote as invasive, DCIS, atypical, proliferative, or non-proliferative. The final diagnostic hierarchical categories for each case were then compared between pathologists, and cases with disagreement were flagged for consensus review.

Figure 2.

(A,B) (H&E): Two of the three expert breast pathologists independently classified this case as ADH, and one classified it as DCIS. During consensus review discussion, there was agreement about the cytological monotony and early architecture of low-grade DCIS, but it was agreed that the architectural changes present were not sufficient for a diagnosis of low-grade DCIS. The consensus diagnosis was ADH. (C) (H&E): Two of three expert breast pathologists independently classified this case as ADH, and one classified it as usual ductal hyperplasia (UDH). Consensus review discussion focused on how the features were borderline between UDH and ADH, and the predictability of their disagreement about subtle features meeting diagnostic criteria. At the consensus conference, it was agreed that there was enough architectural atypia present for a diagnosis of ADH.

Figure 3.

(A,B) (H&E): This lesion was diagnosed as ADH by two of three expert breast pathologists, and coded by research coordinators as 'atypia'. The third pathologist utilized the 'other' category present in the original paper diagnostic forms, and handwrote a diagnosis of 'atypical papillary lesion', which was initially coded as a diagnostic disagreement. Consensus discussions made it clear that the case was miscoded, and these diagnostic coding/study methodology-related errors were greatly reduced by the adoption of an electronic data entry diagnostic form.

Table 1

Themes identified as reasons for diagnostic variability, and their frequency as the root cause of diagnostic variability

Theme	Definition	Frequency ranking
Pathologist-related		
Professional differences of opinion on features meeting diagnostic criteria	Discussion focused on subtle differences of professional opinion about whether the features present met the criteria for a specific diagnosis	1st
Not noting a focal diagnostic finding	Pathologist verbally acknowledged not noting the diagnostic area of the slide (all of these were focal findings)	2nd
Different diagnostic philosophy (clinical impact versus morphology)	Discussion focused on differences in taking into account the potential clinical relevance of a diagnosis versus utilizing strictly morphological features	3rd
Different diagnostic criteria	Discussion focused on pathologists' use of different diagnostic criteria for a specific diagnosis in a given case	4th
Different diagnostic features noted	Discussion focused on disagreement about the specific morphological features present	4th
Diagnostic-coding/study methodology-related		
'Other' category diagnosis miscoded	A handwritten diagnosis in the 'other diagnosis' category was not recognized by coders as a specific diagnosis	1st
Disagreement only based on a change in diagnostic categories of study	Cases where there would have been agreement using the final diagnostic categories	2nd
Not a clinically relevant difference	Discussion focused on differences in benign diagnoses (typically between proliferative and non-proliferative categories)	3rd
Forced into diagnostic category	Pathologists indicated in discussion that they would not have been definitive about a specific diagnosis, or that the specific terminology they would have used was not an option in the diagnostic assessment form	4th
Specimen-related		
Poor slide quality/artefacts	Discussion focused on the challenge of making a diagnosis in the setting of poor histology or obscuring tissue artefacts	1st
Limited diagnostic material	Discussion focused on the challenge of making a diagnosis given the small quantity of tissue with the diagnostic finding.	2nd
Diagnosis required additional stains	Discussion focused on the challenge of making a diagnosis without the use of ancillary studies such as additional levels or immunohistochemistry	2nd

Table 2

Illustrative quotations from consensus meeting transcripts exemplifying reasons for diagnostic variability

Reason for diagnostic variability	Examples from consensus conference transcripts
Pathologist-related	
Professional differences of opinion on features meeting diagnostic criteria	<p>‘This is a continuous biologic spectrum on which we have imposed arbitrary cut points, and it is very difficult to keep that arbitrary cut point in the same space from day to day and from pathologist to pathologist’</p> <p>‘I wish there was a way to note which cases are borderline so it doesn’t look like someone got it dead wrong ... I think there is agreement that it’s a borderline lesion and we’re just falling on different sides of the fence’</p>
Not noting a focal diagnostic finding	‘Looking at it again, we all agree there is focal ADH ... I wonder if this would be a little more obvious on additional levels’
Different diagnostic philosophy (clinical impact versus morphology)	‘To me this could be DCIS in a papilloma. But I wouldn’t want to call that on the core. I’d back down to ADH because that gives me flexibility ... to classify it on final surgical pathology and not to have overdiagnosed it’
Diagnostic coding/study methodology-related	
‘Other’ category diagnosis miscoded	‘I think what we’ve uncovered here is more of a systems issue with regard to worksheets and data entry rather than diagnostic discrepancies’
Not a clinically relevant difference	‘I think what happens is, if it doesn’t matter clinically, we don’t spend a lot of time worrying about the exact diagnosis’
Forced into a diagnostic category	‘... in a study you’re forced to pigeon-hole a diagnosis into a slot. In clinical practice, I think we’d all do the same thing; we’d make an equivocal diagnosis that was not definitive, that was enough to get clinicians to do the right thing’
Specimen-related	
Poor slide quality/artefacts	‘The tissue is poorly preserved ... I mean if you really want to test pathologists in terms of their interpretative abilities, this is more of a judgement call in terms of how you’d manage the case ... rather than how you’d interpret it’