HHS Public Access

Author manuscript

Ann Intern Med. Author manuscript; available in PMC 2017 May 17.

Published in final edited form as:

Ann Intern Med. 2016 May 17; 164(10): 649–655. doi:10.7326/M15-0964.

Variability in Pathologists' Interpretations of Individual Breast Biopsy Slides: A Population Perspective

Joann G. Elmore, MD, MPH, **Heidi D. Nelson, MD, MPH**, **Margaret S. Pepe, PhD**, **Gary M. Longton, MS**, **Anna N.A. Tosteson, ScD**, **Berta Geller, EdD**, **Tracy Onega, PhD**, **Patricia A. Carney, PhD**, **Sara L. Jackson, MD, MPH**, **Kimberly H. Allison, MD**, and **Donald L. Weaver, MD**

University of Washington School of Medicine and Fred Hutchinson Cancer Research Center, Seattle, Washington; Providence Cancer Center, Providence Health & Services Oregon, and Oregon Health & Science University, Portland, Oregon; Geisel School of Medicine at Dartmouth,

Requests for Single Reprints: Joann G. Elmore, MD, MPH, University of Washington, Mailbox 359780, 325 Ninth Avenue, Seattle, WA 98104.

Current Author Addresses: Drs. Elmore and Jackson: University of Washington, Mailbox 359780, 325 Ninth Avenue, Seattle, WA 98104.

Dr. Nelson: Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code BICC, Portland, OR 97239.

Dr. Pepe and Mr. Longton: Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, M2-B500, 1100 Fairview Avenue North, PO Box 19024, Seattle, WA 98109-1024.

Dr. Tosteson: Geisel School of Medicine at Dartmouth, One Medical Center Drive (HB7505), Lebanon, NH 03756.

Dr. Geller: Family Medicine, University of Vermont, 1 South Prospect Street, Burlington, VT 05401.

Dr. Onega: Section of Biostatistics & Epidemiology, Geisel School of Medicine at Dartmouth, One Medical Center Drive (HB7937), Lebanon, NH 03756.

Dr. Carney: Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code FM, Portland, OR 97239.

Dr. Allison: Department of Pathology, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 93195.

Dr. Weaver: Department of Pathology, University of Vermont, Courtyard at Given, 89 Beaumont Avenue, Burlington, VT 05405-0068.

Current author addresses and author contributions are available at www.annals.org.

Author Contributions: Conception and design: J.G. Elmore, H.D. Nelson, M.S. Pepe, A.N.A. Tosteson, B. Geller, P.A. Carney, K.H. Allison, D.L. Weaver.

Analysis and interpretation of the data: J.G. Elmore, H.D. Nelson, M.S. Pepe, G.M. Longton, A.N.A. Tosteson, B. Geller, T. Onega, P.A. Carney, S.L. Jackson, K.H. Allison, D.L. Weaver.

Drafting of the article: J.G. Elmore, H.D. Nelson, M.S. Pepe, A.N.A. Tosteson, B. Geller, T. Onega, P.A. Carney, S.L. Jackson, K.H. Allison, D.L. Weaver.

Critical revision of the article for important intellectual content: J.G. Elmore, H.D. Nelson, M.S. Pepe, G.M. Longton, A.N.A. Tosteson, B. Geller, T. Onega, P.A. Carney, K.H. Allison, D.L. Weaver.

Final approval of the article: J.G. Elmore, H.D. Nelson, M.S. Pepe, G.M. Longton, A.N.A. Tosteson, B. Geller, T. Onega, P.A. Carney, S.L. Jackson, K.H. Allison, D.L. Weaver.

Provision of study materials or patients: J.G. Elmore, T. Onega, P.A. Carney, D.L. Weaver.

Statistical expertise: M.S. Pepe, G.M. Longton, A.N.A. Tosteson.

Obtaining of funding: J.G. Elmore, M.S. Pepe, A.N.A. Tosteson, T. Onega, P.A. Carney, D.L. Weaver.

Administrative, technical, or logistic support: J.G. Elmore, S.L. Jackson, D.L. Weaver.

Collection and assembly of data: J.G. Elmore, H.D. Nelson, B. Geller, T. Onega, P.A. Carney, S.L. Jackson, D.L. Weaver.

Note: Drs. Elmore and Pepe 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.

Disclaimer: The content of this article is solely the responsibility of the authors and does not necessarily represent the views of the National Cancer Institute or the National Institutes of Health.

Disclosures: Dr. Elmore reports grants from the National Cancer Institute during the conduct of the study. Mr. Longton reports grants from the National Cancer Institute during the conduct of the study. Dr. Tosteson reports grants from the National Cancer Institute during the conduct of the study. Dr. Onega reports grants from the National Cancer Institute during the conduct of the study. Dr. Weaver reports grants from the National Cancer Institute during the conduct of the study. Authors not named here have disclosed no conflicts of interest. Disclosures can also be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M15-0964.

Reproducible Research Statement: *Study protocol, statistical code, and data set:* Readers may contact the authors directly to discuss the study protocol, the statistical code, or the data set from which the results were derived (jelmore@u.washington.edu).

Lebanon, New Hampshire; University of Vermont, Burlington, Vermont; and Stanford University School of Medicine, Stanford, California.

Abstract

Background—The effect of physician diagnostic variability on accuracy at a population level depends on the prevalence of diagnoses.

Objective—To estimate how diagnostic variability affects accuracy from the perspective of a U.S. woman aged 50 to 59 years having a breast biopsy.

Design—Applied probability using Bayes theorem.

Setting—B-Path (Breast Pathology) Study comparing pathologists' interpretations of a single biopsy slide versus a reference consensus interpretation from 3 experts.

Participants—115 practicing pathologists (6900 total interpretations from 240 distinct cases).

Measurements—A single representative slide from each of the 240 cases was used to estimate the proportion of biopsies with a diagnosis that would be verified if the same slide were interpreted by a reference group of 3 expert pathologists. Probabilities of confirmation (predictive values) were estimated using B-Path Study results and prevalence of biopsy diagnoses for women aged 50 to 59 years in the Breast Cancer Surveillance Consortium.

Results—Overall, if 1 representative slide were used per case, 92.3% (95% CI, 91.4% to 93.1%) of breast biopsy diagnoses would be verified by reference consensus diagnoses, with 4.6% (CI, 3.9% to 5.3%) overinterpreted and 3.2% (CI, 2.7% to 3.6%) underinterpreted. Verification of invasive breast cancer and benign without atypia diagnoses is highly probable; estimated predictive values were 97.7% (CI, 96.5% to 98.7%) and 97.1% (CI, 96.7% to 97.4%), respectively. Verification is less probable for atypia (53.6% overinterpreted and 8.6% underinterpreted) and ductal carcinoma in situ (DCIS) (18.5% overinterpreted and 11.8% underinterpreted).

Limitations—Estimates are based on a testing situation with 1 slide used per case and without access to second opinions. Population-adjusted estimates may differ for women from other age groups, unscreened women, or women in different practice settings.

Conclusion—This analysis, based on interpretation of a single breast biopsy slide per case, predicts a low likelihood that a diagnosis of atypia or DCIS would be verified by a reference consensus diagnosis. This diagnostic gray zone should be considered in clinical management decisions in patients with these diagnoses.

Primary Funding Source—National Cancer Institute.

Results of the B-Path (Breast Pathology) Study, an evaluation of diagnostic agreement among pathologists interpreting breast biopsy specimens, indicated marked variability across diagnostic categories (1). The B-Path Study and others have reported high agreement for slides interpreted as invasive breast cancer or benign cases without atypia but much lower for those interpreted as ductal carcinoma in situ (DCIS) or atypia (1–3). These results raise concerns that interpretations of breast biopsy specimens in clinical practice may be inaccurate. For example, 1 of 4 breast biopsy assessments in the B-Path Study disagreed with the expert reference consensus diagnosis. This result was highlighted by the media,

with such statements as, “A recent study showed that 25 percent of the time, pathologists disagree with one another in making a diagnosis of cancer” and “as many as one-in-four biopsies are incorrectly diagnosed” (4, 5).

It would be incorrect to infer that the B-Path Study's overall discordance rate of 25%, based on a test set with 1 slide per case, is an estimate of the expected accuracy of breast pathology in general clinical practice. The study included higher proportions of cases of DCIS and atypia than typically seen in clinical practice, and the overall discordance rate was not intended to reflect population impact. Applying the B-Path Study results to patient populations and communicating the results to patients requires additional analyses that account for population-based prevalence rates for breast biopsy outcomes.

The purpose of this analysis was to estimate the effect of variation in the interpretation of breast biopsy specimens from the perspective of a woman having a biopsy, using U.S. population–adjusted estimates derived from the B-Path Study. This approach provides more clinically relevant estimates of accuracy than previously reported unadjusted estimates.

Methods

Overview

We estimated the probability that a pathologist's interpretation of a single diagnostically representative breast biopsy slide would be confirmed by a consensus-based reference standard derived from 3 expert breast pathologists interpreting the same slide. For example, if a single slide from a woman's biopsy is interpreted as DCIS, how likely is she to obtain the same diagnosis if a panel of 3 expert pathologists provides a consensus interpretation of the same slide? We calculated the probabilities (predictive values) using Bayes theorem, combining results from the B-Path Study with published data of the prevalence of breast pathology diagnoses in women aged 50 to 59 years from the Breast Cancer Surveillance Consortium (BCSC) (6). The BCSC is a nationally representative registry of women having mammography in the United States.

The B-Path Study

The B-Path Study invited U.S. pathologists to interpret 1 of 4 test sets of 60 breast biopsy slides (240 total cases, with 1 slide per case). These included 72 benign cases without atypia (24 nonproliferative and 48 proliferative without atypia), 72 with atypia (for example, atypical ductal hyperplasia), 73 with DCIS, and 23 with invasive breast cancer. The proportional representation of these categories differed from population-based prevalence, where most cases are benign without atypia.

Sixty-five percent of invited pathologists who responded were eligible and consented to participate, and 115 completed the study, providing a total of 6900 individual case interpretations. Pathologists were blinded to the interpretations of other study pathologists. Their interpretations were compared with the reference consensus diagnoses, as defined by a panel of 3 experienced breast pathologists who are internationally recognized for research and continuing medical education on diagnostic breast pathology. The 3 panelists reviewed the cases independently and discussed discordant diagnoses by using a multihead

microscope to evaluate the cases until a consensus diagnosis was established for each (7, 8). The reference panel members' concordance with the final consensus diagnoses was 90%, whereas participants' concordance with the same consensus diagnoses was 75%. Unanimous agreement among the 3 reference pathologists' independent diagnoses was 75%, and the average level of unanimous agreement among all possible combinations of 3 participants was 58% (9).

Prevalence of Breast Pathology Diagnoses

Prevalence rates for each diagnostic category were based on BCSC data from women aged 50 to 59 years who had recent screening mammography (25.1% for invasive breast cancer, 6.1% for DCIS, 3.9% for atypia, and 64.9% for benign cases without atypia) (**Appendix**, available at www.annals.org) (6). Prevalence rates by diagnostic category within the B-Path test set and among U.S. women aged 50 to 59 years are shown in **Figure 1**.

Statistical Analysis

The predictive value of a given diagnosis is the probability that the diagnosis is correct. In this study, the diagnosis was based on a single slide interpreted by a participating pathologist, with the reference standard being the consensus diagnosis of the 3 expert pathologists interpreting the same slide. Predictive value estimates were based on results of the 115 pathologists' interpretations of the test set cases and the prevalence of biopsy outcomes in the BCSC (6). Bayes theorem (10) was used to calculate the probability of obtaining a reference consensus diagnosis ("D") given the case interpretation ("T") by a single pathologist, as follows:

$$Prob\,[\mathrm{D}|\mathrm{T}] = (Prob\,[\mathrm{T}|\mathrm{D}] \times Prob\,[\mathrm{D}])\,/Prob\,[\mathrm{T}]$$

Prob[D] is the prevalence of interpretive category "D" (such as DCIS or atypia) among women who have biopsy. Prob[T|D] is the probability that a single slide interpreted by the 3 experts as "D" will be interpreted by a single study pathologist as "T". Prob[T] is equal to Sum (Prob[T|D] × Prob[D]), where "D" represents the reference interpretation and the sum is over all possible interpretive categories for "D". An illustration of the calculation is provided in the **Appendix**. Because the predictive value is the probability that a diagnosis ("t") will be confirmed by the reference consensus diagnosis, it is calculated as Prob(D=t|T=t).

Probabilities of interpretations by pathologists relative to the reference consensus diagnosis (P[T|D]) were derived from the B-Path Study data and are described in **Appendix Tables 1** and **2** (available at www.annals.org). We previously published interpretation rates based on 4 diagnostic categories and used 5 categories in this analysis to further adjust the benign without atypia category using its component outcomes (nonproliferative and proliferative without atypia) (1).

We defined overinterpretation as a diagnosis that was at a higher level of severity than the reference consensus diagnosis and underinterpretation as a diagnosis at a lower level of severity than the reference consensus diagnosis. Overinterpretation rates were 26% for

benign nonproliferative, 18% for proliferative without atypia, 17% for atypia, and 3% for DCIS. Underinterpretation rates were 8% for proliferative without atypia, 35% for atypia, 13% for DCIS, and 4% for invasive breast cancer. Because the test set oversampled proliferative cases relative to nonproliferative cases in the benign without atypia category, we calculated the predictive values for 5 diagnostic categories and then collapsed the proliferative and nonproliferative categories into the benign without atypia interpretations. Bootstrapping of data from the B-Path Study was used to calculate CIs for the predictive values by resampling study pathologists randomly and with replacement. Our CIs do not account for sampling variability in the cases.

Role of the Funding Source

This work was supported by the National Cancer Institute of the National Institutes of Health. The content is solely the responsibility of the authors.

Results

When a single slide is used to represent the breast biopsy, population-adjusted predictive value estimates indicate that confirmation of pathologists' interpretations by the expert reference consensus panel would occur in 92.3% (95% CI, 91.4% to 93.1%) of biopsies overall, with 4.6% (CI, 3.9% to 5.3%) overinterpreted and 3.2% (CI, 2.7% to 3.6%) underinterpreted. These estimates assume that the representative diagnostic features of the case are present on the slide examined and do not account for the effect of second opinions that might be obtained in clinical practice.

As noted in **Figure 1**, most women having breast biopsy in U.S. clinical practice receive a diagnosis of benign without atypia. For these women, our analysis indicated that diagnostic agreement with the reference panel would be high (97.1% [CI, 96.7% to 97.4%]) (**Table 1** and **Figures 2** and **3**). Only 2.1% (CI, 1.9% to 2.4%) of the biopsy slides would be interpreted at the higher diagnostic category of atypia by the reference consensus panel, with fewer than 1% interpreted as DCIS (0.6% [CI, 0.5% to 0.7%]) or invasive breast cancer (0.2% [CI, 0.0% to 0.4%]).

Most diagnoses of atypia on a single slide would be overinterpretations by the pathologist; the reference consensus panel would interpret 53.6% (CI, 47.9% to 58.3%) as benign without atypia and 8.6% (CI, 7.0% to 10.5%) as DCIS. The reference panel noted that these DCIS cases would likely be low-grade rather than high-grade DCIS.

For the cases interpreted as DCIS, the reference consensus panel would interpret 9.5% (CI, 5.7% to 13.6%) as benign without atypia, 9.0% (CI, 7.8% to 10.2%) as atypia, and 11.8% (CI, 7.6% to 15.7%) as invasive breast cancer. The last estimate may have been influenced by the presence of 1 case of DCIS with focal microinvasion on the slide that was difficult to identify and was frequently diagnosed as DCIS by study pathologists. The reference panel noted that this microinvasive focus would not lead to a significant change in treatment or outcome.

Finally, for women receiving a single-slide interpretation of invasive breast cancer, diagnostic agreement with the reference consensus panel would be 97.7% (CI, 96.5% to 98.7%).

In practice, pathologists often obtain second opinions, and diagnoses that are on the borderline between 2 categories might be factored into treatment decisions. Therefore, diagnostic agreement with the reference consensus panel is shown by whether the pathologist noted that the case was or was not borderline and whether he or she desired a second opinion (**Table 2**). Diagnostic agreement with the reference consensus panel for atypia was less than 50% regardless of the pathologists' desire for a second opinion or whether they noted that the case was borderline. When we restricted the analysis to slides for which pathologists did not consider the case borderline, probabilities of confirmation by the reference consensus panel were 36.8% for atypia and 76.3% for DCIS. When the pathologists did not want a second opinion on the slide, the probability of confirmation was 78.1% for DCIS and 42.5% for atypia.

Discussion

The B-Path Study showed high diagnostic agreement between pathologists and a reference consensus panel of 3 expert breast pathologists for invasive breast cancer but substantially lower agreement for interpretations of DCIS and especially atypia (1). To extrapolate the B-Path Study results to estimates more relevant to clinical practice, the current analysis included adjustments for prevalence of outcomes in a mammography screening population of women aged 50 to 59 years in the United States. Our results, based on the use of 1 slide per case, suggest that more than 92% of interpretations of breast biopsy specimens in this group of women would be likely to agree with the interpretations of the reference consensus panel. Actual accuracy may be higher due to the effects of obtaining second opinions and context (evaluating >1 slide or special stains).

Although the prevalence of atypia and DCIS diagnoses is low among the total breast biopsies performed each year, the markedly lower diagnostic agreement rates for these categories should not be overlooked or minimized. These noninvasive but potentially high-risk breast lesions represent a gray area with subjective boundaries imposed on a biological continuum; there is not always a "right" or "wrong" diagnosis and, as in many areas of medicine, professional opinions may differ. For women having breast biopsy, our results suggest that nearly 1 in 5 (18.5%) with a diagnosis of DCIS would have her biopsy specimen interpreted as atypia or benign by our reference consensus panel (with the limiting assumptions that the diagnostic features are present on a single slide and no second opinions are obtained).

Overdiagnosis of DCIS has recently been discussed in the literature (11–14). The expressed concern is related to increased detection resulting from widespread use of screening mammography. Our results suggest that overinterpretation of the pathologic findings may contribute to overdiagnosis and overtreatment of DCIS. In current practice, most women diagnosed with DCIS are offered lumpectomy and radiation therapy or total mastectomy, and they may also be offered adjuvant hormonal therapy for 5 to 10 years to reduce

recurrence risk. Most of the diagnostic variability in pathology is likely due to differentiating atypia from low-grade DCIS, a diagnostic challenge that may be due to imposing categorical diagnoses on a biological continuum of disease and may not necessarily reflect the accuracy of the pathologist.

We also estimated that slightly more than half of breast biopsies with diagnoses of atypia (53.6%) based on a single representative slide would be interpreted as benign cases without atypia by our reference panel. Although diagnostic variability for atypia and DCIS has been noted as a particular challenge for nearly 25 years (2, 3), our prevalence-adjusted estimates shed new light on potential diagnostic trends at a population level and suggest that higher levels of overinterpretation rather than underinterpretation may occur in practice. The population prevalence of atypia has been estimated to range between 3.9% and 10% of all breast biopsy interpretations in the United States (6, 15), which equates to a large number of women affected annually given the high number of breast biopsies performed each year. A diagnosis of atypia on a core biopsy has significant implications; atypical ductal hyperplasia has been associated with risk for concurrent and future cancer, and a diagnosis of atypia on a core biopsy is generally followed by an excision biopsy, placement in a high-risk screening group, and consideration for risk-reducing hormonal therapy. Thus, overdiagnosis of atypia may lead to unnecessary surgery, follow-up, and treatment (16, 17).

The practice of surgical pathology involves systematic sampling of tissue specimens; evaluation of multiple slides; and dynamic opportunities to evaluate additional tissue from paraffin blocks, obtain immunohistochemical or molecular markers, and consult colleagues on challenging cases. The B-Path Study did not evaluate the complete diagnostic pathway but focused on the pathologist first reviewing a case. Thus, the results help to define knowledge gaps on which systemic quality improvements can be built.

Studies of physician diagnostic concordance are challenging to design and implement; perfect simulation of the practice of medicine is rarely possible. The underlying data should be evaluated in context, and their limitations should be considered. In addition to having augmented test cases of atypia and DCIS, the B-Path Study provided data from a testing situation in which pathologists gave interpretations based on only 1 slide per case and were not given the opportunity to obtain additional clinical history, additional testing, or a second opinion from a colleague. However, even in clinical practice, a biopsy diagnosis and recommendation can hinge on a single focus on a single slide, and this was the premise presented to the participating pathologists: Assume that the most diagnostic area for the case is present on the test set slide.

In clinical practice, pathologists are able to obtain second opinions and indicate when they consider diagnoses to be borderline. Participants were asked to record whether a case was on the borderline between 2 diagnoses and to indicate whether they would obtain a second consultative opinion. These data provide additional insight into current practice and opportunities for diagnostic improvement. We calculated the predictive value for interpretations for which pathologists would not desire second opinions and found concerning levels of disagreement with the reference consensus panel for slides showing atypia and DCIS. Similarly, slides that were not considered borderline diagnoses of atypia

and DCIS also had high probabilities of not being confirmed by the expert reference consensus panel. Pathologists had higher rates of desiring second opinions or noting a case was borderline when they were less likely to agree with the reference consensus diagnosis, but this was not the case for atypia. Pathologists' diagnoses of atypia had markedly low agreement with the reference standard for all cases, regardless of whether the diagnosis was noted as borderline or the pathologist desired a second opinion.

The B-Path Study reference standard was defined as the consensus diagnosis of 3 experienced breast pathologists. Although the consensus diagnosis may be a reasonable reference standard for a scientific study (9) and, from a patient's perspective, obtaining opinions from 3 experienced pathologists on a case might seem ideal, there is no guarantee that this reference standard represents biological truth. In addition, as stated earlier, differences in diagnostic opinion between 2 or more pathologists may reflect the underlying biological uncertainty inherent in a particular case rather than the diagnostic accuracy of the pathologist. Additional research is needed to determine whether objective measures of diagnostic uncertainty could be integrated into management of breast disease rather than expecting pathologists to always make a definitive microscopic diagnosis.

Our results are based on the limited number of B-Path cases, and histologic data from other populations might differ. We should also note that we used diagnostic prevalence rates based on a population of women in their 50s who were screened using film mammography, and results may differ with newer technologies, such as digital mammography, magnetic resonance imaging, and tomosynthesis. In addition, our estimates may not reflect outcomes for women in other age groups, unscreened women, or women from different countries. Calculations for other populations can be performed by following the example in the *Appendix* and substituting alternative population-specific prevalence estimates of biopsy outcomes.

In summary, we estimate that the initial interpretation of a single breast biopsy slide with representative diagnostic features would disagree with a reference consensus diagnosis for about 8% of women having a biopsy, with more overinterpretations than underinterpretations among the discordant cases. Of note, more than 97% of interpretations of invasive breast cancer and benign cases without atypia from a single slide would agree with the reference diagnosis, but reference diagnosis verification for DCIS and especially atypia is predicted to be much lower. Efforts to reduce diagnostic variability need to be considered and evaluated and might include educational programs, improved diagnostic techniques, or second-review policies. Alternatively, women with borderline breast lesions that are difficult to categorize, such as atypical ductal hyperplasia and low-grade DCIS, may benefit from revised guidelines for clinical treatment and management given the degree of diagnostic variability and biological overlap between these diagnostic categories.

Acknowledgment

The authors appreciate the efforts of the pathologists who participated in this study.

Financial Support: This work was supported by the National Cancer Institute of the National Institutes of Health (award numbers R01 CA140560, U01CA86082, U01 CA70013, and R01 CA172343) and by the National Cancer Institute–funded BCSC (award number HHSN261201100031C). The collection of cancer and vital status data used

in this study was supported in part by several state public health departments and cancer registries throughout the United States. A full description of sources is available at www.breastscreening.cancer.gov/work/acknowledgement.html.

Appendix: Calculating Frequencies of Reference Standard Diagnostic Categories for Women in the Population With Biopsy Slides Interpreted in Each Category by Single Pathologists

The calculations presented here illustrate the use of Bayes theorem that is described in the Methods section of the article. This illustration will allow readers to apply the B-Path results to other populations as long as information about the prevalence of true diagnostic outcomes is available. The B-Path Study used a test set of slides with overrepresentation of cases in several reference diagnostic categories. For cases in each category according to the reference consensus diagnosis, the cross-classified data from B-Path (**Appendix Table 1**) allowed us to calculate the frequency of interpretations by study pathologists in each category. These are shown as the column percentages in **Appendix Table 2**. For example, 17.1% of cases classified as atypia by the expert reference consensus panel were classified as DCIS by study pathologists.

For women aged 50 to 59 years having screening mammography, the proportions in each diagnostic category are 0.516 (benign nonproliferative), 0.133 (benign proliferative), 0.039 (atypia), 0.061 (DCIS), and 0.251 (invasive breast cancer). The bottom row of **Appendix Table 3** shows the expected numbers in each category for 1000 random women. The columns of **Appendix Table 3** show how the number in each category would likely be distributed according to pathologist interpretations. These entries were calculated by multiplying the column total by the column percentages from **Appendix Table 2**.

We next combined the 2 benign categories to arrive at the 4 × 4 distribution of classifications for these 1000 women in **Appendix Table 4**.

Appendix Table 4 also shows the totals in each row. Using these row totals as denominators, in **Appendix Table 5** we calculated the percentages that would be in each diagnostic category according to the reference consensus diagnosis. For example, of the 629.8 pathologist interpretations in the benign category, 611 (97.0%) would be classified as benign and 13.5 (2.1%) would be classified as atypia by the reference consensus panel. The entries in **Appendix Table 5** are the probabilities shown in **Table 1**. There are some discrepancies with the entries in **Table 1** in the first decimal place. These are due to rounding in our illustrative calculations; **Table 1** has the more precise calculations.

Appendix Table 1

Cross-Classification of Study Pathologist Interpretations of Biopsy Slides by Expert Reference Interpretations of the Same Slides in the Breast Pathology Study *

Pathologist Interpretation	Reference Diagnosis, *n*					Total, *n*
	Benign Nonproliferative	Benign Proliferative	Atypia	DCIS	Invasive Breast Cancer	
Benign nonproliferative	511	112	81	67	1	772
Benign proliferative	161	1019	638	66	2	1886
Atypia	11	189	990	146	0	1336
DCIS	4	42	353	1764	23	2186
Invasive breast cancer	3	18	8	54	637	720
Total	690	1380	2070	2097	663	6900

DCIS = ductal carcinoma in situ.

* The study used a test set of cases with overrepresentation of more difficult diagnostic categories.

Appendix Table 2

Interpretation Rates Observed in the Breast Pathology Study

Pathologist Interpretation	Reference Diagnosis, %				
	Benign Nonproliferative	Benign Proliferative	Atypia	DCIS	Invasive Breast Cancer
Benign nonproliferative	74.0	8.1	3.9	3.2	0.2
Benign proliferative	23.3	73.8	30.8	3.2	0.3
Atypia	1.6	13.7	47.8	7.0	0.0
DCIS	0.6	3.0	17.1	84.1	3.5
Invasive breast cancer	0.4	1.3	0.4	2.6	96.1
Total	100.0	100.0	100.0	100.0	100.0

DCIS = ductal carcinoma in situ.

Appendix Table 3

Expected Counts for 1000 Women in the U.S. Mammography Screening Population Aged 50-59 y

Variable	Reference Diagnosis, *n*					Total, *n*
	Benign Nonproliferative	Benign Proliferative	Atypia	DCIS	Invasive Breast Cancer	
Pathologist interpretation						
Benign nonproliferative	381.8	10.8	1.5	2.0	0.5	-
Benign proliferative	120.2	98.2	12.0	2.0	0.8	-
Atypia	8.3	18.2	18.6	4.3	0.0	-
DCIS	3.1	4.0	6.7	51.3	8.8	-
Invasive breast cancer	2.1	1.7	0.2	1.6	241.2	-
Total women, *n* *	516	133	39	61	251	1000

DCIS = ductal carcinoma in situ.

* Expected totals according to population prevalence estimates (6).

Appendix Table 4

Expected Counts for 1000 Women From the U.S. Mammography Screening Population Aged 50-59 y: Benign Nonproliferative and Benign Proliferative Combined

Variable	Reference Diagnosis, *n*				Total, *n*
	Benign	Atypia	DCIS	Invasive Breast Cancer	
Pathologist interpretation					
Benign*	611	13.5	4.0	1.3	629.8
Atypia	26.5	18.6	4.3	0.0	49.4
DCIS	7.1	6.7	51.3	8.8	73.9
Invasive breast cancer	3.8	0.2	1.6	241.2	246.8
Total women, *n*	-	-	-	-	1000

DCIS = ductal carcinoma in situ.

*Benign nonproliferative and benign proliferative values from **Appendix Table 3** are combined.

Appendix Table 5

Estimated Frequency of Reference Consensus for Randomly Selected Cases From the U.S. Mammography Screening Population Aged 50-59 y With 1 Slide Interpreted by a Single Pathologist*

Pathologist Interpretation	Reference Diagnosis, %				Total, %
	Benign	Atypia	DCIS	Invasive Breast Cancer	
Benign	97.0	2.1	0.6	0.2	100
Atypia	53.6	37.7	8.7	0	100
DCIS	9.6	9.1	69.4	11.9	100
Invasive breast cancer	1.5	0.1	0.6	97.7	100

DCIS = ductal carcinoma in situ.

*Values are the probabilities from **Table 1** presented as percentages. These calculations can be applied to other populations by substituting the appropriate population prevalence estimates in the bottom row of **Appendix Table 3**.

References

1. Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson AN, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA. 2015; 313:1122–32. [PMID: 25781441] doi:10.1001/jama.2015.1405. [PubMed: 25781441]
2. Rosai J. Borderline epithelial lesions of the breast. Am J Surg Pathol. 1991; 15:209–21. [PMID: 1847606]. [PubMed: 1847606]
3. Schnitt SJ, Connolly JL, Tavassoli FA, Fechner RE, Kempson RL, Gelman R, et al. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. Am J Surg Pathol. 1992; 16:1133–43. [PMID: 1463092]. [PubMed: 1463092]
4. Holt, L.; Thompson, A. Study finds breast cancer biopsies are often misdiagnosed. NBC; Mar 17. 2015 at www.nbcnews.com/nightly-news/video/study-finds-breast-cancer-biopsies-areoften-misdiagnosed-414596163784 [11 September 2015]
5. Abraham, J. [11 September 2015] Getting a second opinion may save more than your breast.. U.S. News & World Report. Jul 27. 2015 at http://health.usnews.com/health-news/patient-advice/articles/2015/07/27/getting-a-second-opinion-may-save-more-than-your-breast

6. Weaver DL, Rosenberg RD, Barlow WE, Ichikawa L, Carney PA, Kerlikowske K, et al. Pathologic findings from the Breast Cancer Surveillance Consortium: population-based outcomes in women undergoing biopsy after screening mammography. Cancer. 2006; 106:732–42. [PMID: 16411214]. [PubMed: 16411214]
7. Allison KH, Reisch LM, Carney PA, Weaver DL, Schnitt SJ, O'Malley FP, et al. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. Histopathology. 2014; 65:240–51. [PMID: 24511905] doi:10.1111/his .12387. [PubMed: 24511905]
8. Oster NV, Carney PA, Allison KH, Weaver DL, Reisch LM, Longton G, 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 Womens Health. 2013; 13:3. [PMID: 23379630] doi:10.1186/1472-6874-13-3. [PubMed: 23379630]
9. Elmore JG, Pepe MS, Weaver DL. Discordant interpretations of breast biopsy specimens by pathologists—reply [Letter]. JAMA. 2015; 314:83–4. [PMID: 26151274] doi:10.1001/jama. 2015.6239. [PubMed: 26151274]
10. Fisher, L.; Van Belle, G. Biostatistics: A Methodology for the Health Sciences. Wiley; New York: 1993.
11. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. N Engl J Med. 2012; 367:1998–2005. [PMID: 23171096] doi:10.1056/ NEJMoa1206809. [PubMed: 23171096]
12. Hall FM. Identification, biopsy, and treatment of poorly understood premalignant, in situ, and indolent low-grade cancers: are we becoming victims of our own success? Radiology. 2010; 254:655–9. [PMID: 20177083] doi:10.1148/radiol.09092100. [PubMed: 20177083]
13. Esserman L, Shieh Y, Thompson I. Rethinking screening for breast cancer and prostate cancer. JAMA. 2009; 302:1685–92. [PMID: 19843904] doi:10.1001/jama.2009.1498. [PubMed: 19843904]
14. Esserman LJ, Thompson IM Jr, Reid B. Overdiagnosis and over-treatment in cancer: an opportunity for improvement. JAMA. 2013; 310:797–8. [PMID: 23896967]. [PubMed: 23896967]
15. Rubin E, Visscher DW, Alexander RW, Urist MM, Maddox WA. Proliferative disease and atypia in biopsies performed for nonpalpable lesions detected mammographically. Cancer. 1988; 61:2077–82. [PMID: 3359405]. [PubMed: 3359405]
16. National Comprehensive Cancer Network. Clinical Practice Guidelines: Breast Cancer Screening and Diagnosis. National Comprehensive Cancer Network; Fort Washington, PA: 2015.
17. Hartmann LC, Degnim AC, Santen RJ, Dupont WD, Ghosh K. Atypical hyperplasia of the breast —risk assessment and management options. N Engl J Med. 2015; 372:78–89. [PMID: 25551530] doi:10 .1056/NEJMsr1407164. [PubMed: 25551530]

EDITORS' NOTES

Context

Variability in the interpretation of breast biopsy slides has been documented, but its effect at a population level depends on the prevalence of diagnoses.

Contribution

To estimate how diagnostic variability affects accuracy from the perspective of a U.S. woman aged 50 to 59 years having a breast biopsy, researchers compared pathologists' interpretations of a single case slide with a reference consensus interpretation from 3 experts. The likelihood that a diagnosis of atypia or ductal carcinoma in situ (DCIS) would be verified by the reference consensus diagnosis was low.

Caution

Pathologists reviewed a single slide.

Implication

Clinicians and patients should be aware that atypia and DCIS represent diagnostic gray zones.

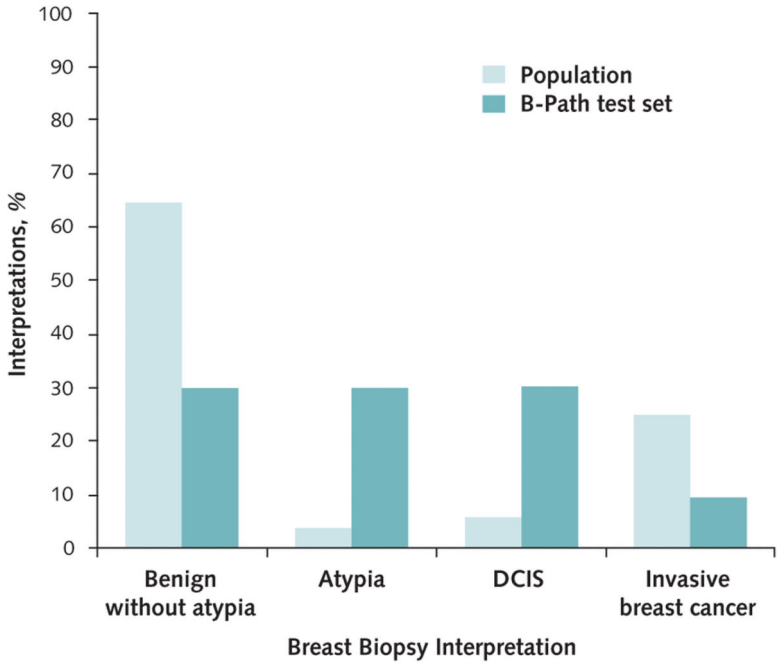

Figure 1.

Prevalence of breast biopsy diagnostic interpretations in the B-Path Study test set and among women aged 50 to 59 y having screening mammography in the United States.

Estimates of population prevalence are from the Breast Cancer Surveillance Consortium (6).

B-Path = Breast Pathology; DCIS = ductal carcinoma in situ.

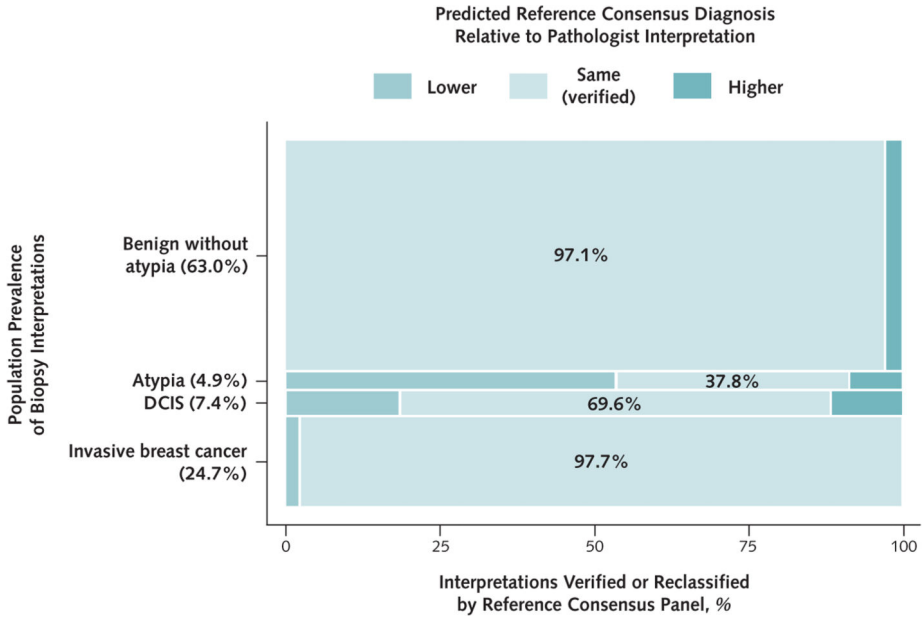

Figure 2.

Predicted proportions of breast biopsy interpretations (based on a single slide per case) that would be verified by the reference consensus panel interpretation or would be classified as overinterpretations or underinterpretations.

DCIS = ductal carcinoma in situ.

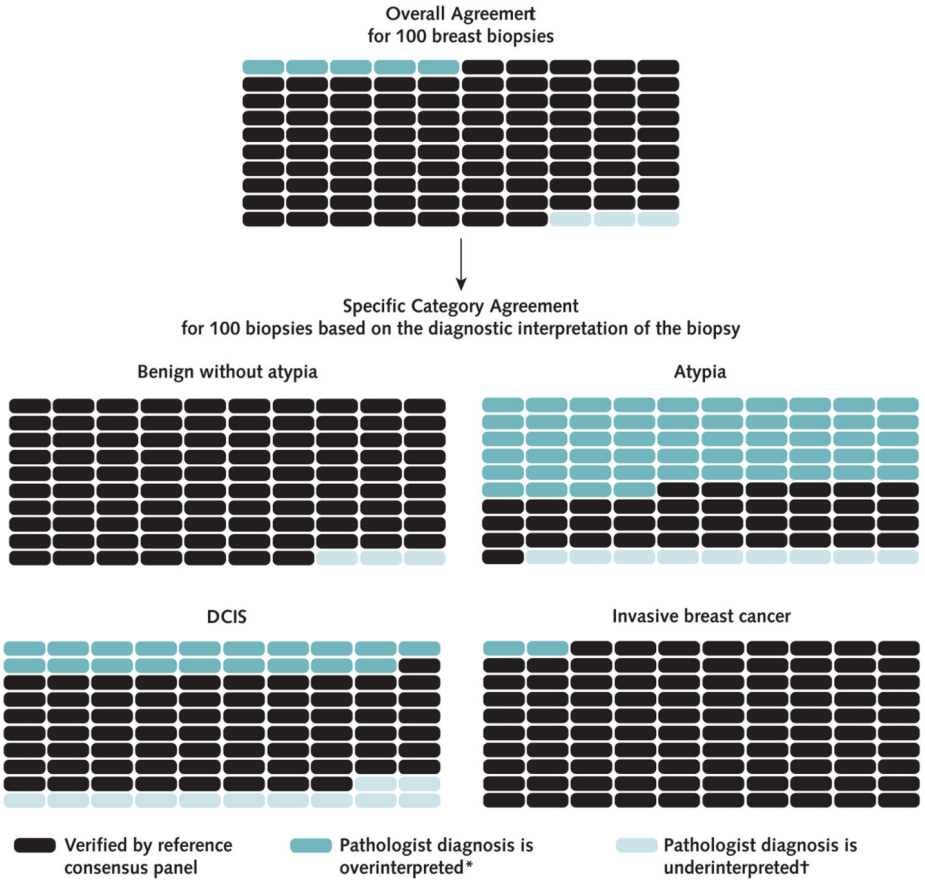

Figure 3.

Predicted outcomes per 100 breast biopsies, overall and by diagnostic category.

Each 100-slide grid demonstrates the predicted number of cases overinterpreted or underinterpreted relative to the reference consensus panel diagnosis, or verified. DCIS = ductal carcinoma in situ.

* Interpretation is more severe than the consensus reference panel.

† Interpretation is less severe than the consensus reference panel.

Table 1

Probability That a Pathologist's Interpretation of a Single-Slide Breast Biopsy Specimen Will Be Verified by the Reference Consensus Interpretation in the U.S. Population of Women Aged 50-59 y Having Screening Mammography

Pathologist Interpretation	Probability of Reference Consensus Interpretation (95% CI), %*				Total, %
	Benign Without Atypia	Atypia	DCIS	Invasive Breast Cancer	
Benign without atypia	**97.1 (96.7-97.4)**	2.1 (1.9-2.4)	0.6 (0.5-0.7)	0.2 (0.0-0.4)	100
Atypia	53.6 (47.9-58.3)	**37.8 (33.6-42.7)**	8.6 (7.0-10.5)	0.0 (0.0-0.0)	100
DCIS	9.5 (5.7-13.6)	9.0 (7.8-10.2)	**69.6 (64.4-75.3)**	11.8 (7.6-15.7)†	100
Invasive breast cancer	1.6 (0.7-2.7)	0.1 (0.0-0.1)	0.6 (0.4-0.9)	**97.7 (96.5-98.7)**	**100**

DCIS = ductal carcinoma in situ.

*Boldface values indicate probabilities of verification by the reference consensus interpretation (i.e., predictive values).

†This estimate may have been influenced by 1 case of DCIS with focal microinvasion that was difficult to identify and was frequently diagnosed as DCIS by study participants. The reference panel noted that this microinvasive focus would not significantly change the treatment or outcome.

Table 2

Probabilities of Verification of a Pathologist's Interpretation by the Reference Consensus Interpretation, Stratified by Whether Cases Were Considered Borderline and Whether Second Opinions Were Desired

Pathologist Interpretation	Probability of Reference Consensus Interpretation (95% CI), %			
	Case Considered Borderline		Second Opinion Desired	
	Yes	No	Yes	No
Benign without atypia	91.3 (89.4-92.9)	97.7 (97.2-98.0)	91.1 (88.7-92.8)	98.1 (97.7-98.4)
Atypia	38.7 (32.7-45.8)	36.8 (30.7-44.2)	35.7 (30.5-41.9)	42.5 (35.2-51.1)
DCIS	46.9 (38.5-56.8)	76.3 (70.0-82.8)	52.5 (44.6-61.7)	78.1 (72.1-84.4)
Invasive breast cancer	63.9 (44.5-81.7)	99.2 (98.8-99.6)	89.4 (82.0-95.1)	99.3 (98.9-99.6)

DCIS = ductal carcinoma in situ.