Evaluation of 12 strategies for obtaining second opinions to improve interpretation of breast histopathology: simulation study

Joann G Elmore,1 Anna NA Tosteson,2,3 Margaret S Pepe,4 Gary M Longton,5 Heidi D Nelson,6 Berta Geller,7 Patricia A Carney,8 Tracy Onega,9 Kimberly H Allison,10 Sara L Jackson,11 Donald L Weaver12

ABSTRACT

OBJECTIVE
To evaluate the potential effect of second opinions on improving the accuracy of diagnostic interpretation of breast histopathology.

DESIGN
Simulation study.

SETTING
12 different strategies for acquiring independent second opinions.

PARTICIPANTS
Interpretations of 240 breast biopsy specimens by 115 pathologists, one slide for each case, compared with reference diagnoses derived by expert consensus.

MAIN OUTCOME MEASURES
Misclassification rates for individual pathologists and for 12 simulated strategies for second opinions. Simulations compared accuracy of diagnoses from single pathologists with that of diagnoses based on pairing interpretations from first and second independent pathologists, where resolution of disagreements was by an independent third pathologist. 12 strategies were evaluated in which acquisition of second opinions depended on initial diagnoses, assessment of case difficulty or borderline characteristics, pathologists’ clinical volumes, or whether a second opinion was required by policy or desired by the pathologists. The 240 cases included benign without atypia (10% non-proliferative, 20% proliferative without atypia), atypia (30%), ductal carcinoma in situ (DCIS, 30%), and invasive cancer (10%). Overall misclassification rates and agreement statistics depended on the composition of the test set, which included a higher prevalence of difficult cases than in typical practice.

RESULTS
Misclassification rates significantly decreased (P<0.001) with all second opinion strategies except for the strategy limiting second opinions only to cases of invasive cancer. The overall misclassification rate decreased from 24.7% to 18.1% when all cases received second opinions (P<0.001). Obtaining both first and second opinions from pathologists with a high volume (≥10 breast biopsy specimens weekly) resulted in the lowest misclassification rate in this test set (14.3%, 95% confidence interval 10.9% to 18.0%). Obtaining second opinions only for cases with initial interpretations of atypia, DCIS, or invasive cancer decreased the over-interpretation of benign cases without atypia from 12.9% to 6.0%. Atypia cases had the highest misclassification rate after single interpretation (52.2%), remaining at more than 34% in all second opinion scenarios.

CONCLUSION
Second opinions can statistically significantly improve diagnostic agreement for pathologists’ interpretations of breast biopsy specimens; however, variability in diagnosis will not be completely eliminated, especially for breast specimens with atypia.

Introduction
Attention to diagnostic errors in the medical literature and mass media has led many to consider obtaining second opinions to prevent errors and improve quality.1,3 For example, obtaining second opinions, such as double reading of screening mammograms, has been associated with improved detection rates for cancer.4,5 Obtaining a second opinion is a strategy commonly suggested to improve diagnostic accuracy in breast disease.6,8 Interpretation of breast pathology is notoriously difficult and rates of disagreement between pathologists are high, especially in cases of atypia (eg, atypical ductal hyperplasia, ADH) and ductal carcinoma in situ (DCIS).6–9 A survey of laboratories in the United States noted that 6.6% of all histopathology cases were reviewed before sign out, suggesting second opinions are often obtained in clinical practice, especially in challenging areas such as breast disease.12 Guidelines have also been published for obtaining second opinions in pathology to prevent medical errors,13 and approximately two thirds of US pathology laboratories have policies, with most requiring a second review
of new diagnoses of invasive cancer. Criteria for when and how to obtain second opinions in breast disease, however, vary considerably.

In our previous study of 252 pathologists, 81% reported requesting second opinions in the absence of institutional policy for at least some of their breast samples, and 96% believed that second opinions improved their diagnostic accuracy. Other studies also suggest possible improvements in patient outcomes. For example, in one study, a second review of 405 node negative cases of breast cancer resulted in substantial modifications in treatments, potentially decreasing unnecessary interventions and subsequent costs. Despite this strong endorsement by practicing pathologists, and multiple studies noting that more than 10% of breast biopsy specimens have important changes in histopathology diagnoses after review, the best guidelines for when to obtain second opinions in pathology are unknown. Comparing the impact of different strategies for obtaining a second opinion on accuracy needs to be evaluated. Many potential strategies exist, ranging from review of every biopsy specimen by multiple pathologists, to obtaining second opinions only for specific case categories (e.g., cases interpreted initially as invasive breast cancer) or only from different pathology situations (e.g., pathologists who see a high volume of cases weekly (10 or more) versus a low volume).

We compared the effect of different criteria for triggering procurement of second opinions on the accuracy of interpretation of breast disease. We designed our study to assess improvements in accuracy in a controlled test situation using data from 6900 individual interpretations by 115 pathologists. We evaluated 12 strategies with different criteria for obtaining second opinions and compared how each approach may affect over-interpretation and under-interpretation rates relative to reference diagnoses.

Methods
Test set cases and consensus reference diagnoses
We used data from the Breast Pathology Study (B-Path), a national study on the accuracy of interpretation of breast tissue. The 240 biopsy cases were divided into four test sets of 60 cases. Breast biopsy specimens were selected from two state registries (NH, VT), which are part of the Breast Cancer Surveillance Consortium sponsored by the National Cancer Institute. Case selection was stratified by age (49% aged 40–49 years, 51% aged ≥50 years), breast density (51% with heterogeneously or extremely dense breast tissue based on mammography findings), and biopsy type (58% core needle, 42% excisional). Three pathologists experienced in breast disease interpreted each case independently before arriving at a consensus reference diagnosis using a modified Delphi approach.

Their diagnoses were categorized using the breast pathology assessment tool and hierarchy for diagnosis (BPATH-Dx). This tool incorporated 14 distinct diagnostic assessments into four main BPATH-Dx categories: benign without atypia (including non-proliferative and proliferative without atypia), atypia (eg, atypical ductal hyperplasia), DCIS, and invasive carcinoma.

Definitions of interpretations and second opinion strategies
The results of single (initial) interpretations, based on the categorical interpretation by each participating pathologist of each case, have been reported previously. We defined interpretations that incorporated second opinions by considering each possible pair of pathologists interpreting the same case and, when disagreement occurred, included resolution by using a third, independent interpretation. Resolution was defined by assigning the case to the BPATH-Dx diagnosis category identified by two of the three pathologists or, if all three disagreed, assigning the middle diagnosis (fig 1).

Patient involvement
No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for recruitment, design, or implementation of the study. No patients were asked to advise on interpretation or writing up of results. There are no plans to disseminate the results of the research to the relevant patient community.

Participating pathologists
Pathologists from eight US states (Alaska, Maine, Minnesota, New Hampshire, New Mexico, Oregon, Vermont, and Washington) were invited to participate. Details of their identification and recruitment have been described elsewhere. Participants were eligible if they had interpreted breast biopsy specimens in the past year, planned to continue for the next year, and were not residents or fellows in training. A web based survey queried participants about personal characteristics, clinical practices, and interpretive experience.

We randomly assigned pathologists to independently interpret one of the four test sets of 60 cases (each case was represented by a single glass slide), and they recorded their interpretations using the online BPATH-Dx tool. Participants also indicated whether the case was borderline between two diagnoses and whether they would obtain a second opinion in their usual clinical practice because of laboratory policies, because it was personally desired, or both. A six point Likert scale was used to assess each case for perceived diagnostic difficulty, with results summarized as a binary variable (difficult cases rated as 4, 5, or 6).
First pathologist’s interpretation

Is the criterion for obtaining a second opinion met?

No

Final interpretation is first pathologist’s interpretation

Yes

Second pathologist’s interpretation

Does second interpretation agree with first?

No

Third pathologist’s interpretation

Final interpretation is common interpretation of first and second pathologists

Yes

Final interpretation is majority interpretation if 2 of 3 pathologists agree or middle interpretation if all disagree

Box 1: Summary of strategies for obtaining second opinions

1. Second opinion applied to all breast biopsy specimens
2. Second opinion obtained only if initial interpretation is atypia, ductal carcinoma in situ (DCIS), or invasive carcinoma
3. Second opinion obtained only if initial interpretation is DCIS or invasive carcinoma
4. Second opinion obtained only if initial interpretation is invasive carcinoma
5. Second opinion obtained only if initial interpretation considered borderline
6. Second opinion obtained only if initial interpretation considered difficult
7. Second opinion obtained only if initial pathologist desired a second opinion
8. Second opinion obtained only if initial diagnosis would prompt a second opinion
9. Second opinion obtained only if initial pathologist desired a second opinion or if required by policy
10. Evaluation restricted to interpretations from initial pathologists who had less clinical experience in breast pathology as they report a low volume of cases weekly (<10); second opinion obtained on all cases from another low volume pathologist, with third brought in for any disagreement from a high volume pathologist
11. Evaluation restricted to interpretations from initial low volume pathologists: second opinion obtained on all cases, with second and third opinion from high volume pathologists
12. Evaluation restricted to interpretations from initial pathologists with high volumes: second opinion obtained on all cases, with second and third opinions from high volume pathologists

Fig 1 | Algorithm for determination of final biopsy interpretation used in the evaluation of different second opinion policy strategies. Up to three pathologists may be needed to obtain a final interpretation. Data are comprised of 5,145,480 observations each involving three independent pathologist interpretations of a single slide from a breast biopsy specimen and are derived from 115 single pathologists interpreting 60 cases each in four test sets

We evaluated 12 strategies for obtaining a second opinion (see box 1), beginning with the strategy where all cases received second opinions. We then evaluated eight selective strategies for obtaining a second opinion, which were determined by criteria based on the initial pathologist’s diagnosis (that is, second opinions were obtained only for cases initially diagnosed as atypia, DCIS, or invasive; DCIS or invasive; or invasive only), determined by the initial pathologist’s assessment of the case (that is, only for cases marked borderline or considered difficult), and determined by whether a second opinion would be required by policy or desired by the pathologist (that is, only for cases where second opinions were required; desired; required or desired).

Finally, we assessed how the clinical volume of the interpreting pathologist affected diagnoses in three strategies with designs shaped by previous findings from the B-Path study. These strategies included combinations of low volume and high volume pathologists providing first, second, and third opinions, as needed, to resolve discordant diagnoses. We defined low to average volume as dealing with fewer than 10 women with breast biopsies per week and high volume as dealing with 10 biopsies or more per week.

Statistical analyses
We assessed rates of over-interpretation, under-interpretation, and overall misclassification compared with the expert consensus diagnosis as reference. Over-interpretation was defined as cases classified by participants at a hierarchically more severe diagnostic BPATH-Dx category compared with the reference diagnosis category, under-interpretation was defined as cases classified lower than the reference diagnosis category, and misclassification was defined as cases either over-interpreted or under-interpreted compared with the reference diagnosis category.

To simulate interpretations that involved obtaining second opinions, we combined the independent interpretations of the study pathologists. For each case, we created an ordered data record of interpretations for every three pathologists who interpreted the case and used the majority or middle interpretation as their final assessment (fig 1). This is analytically equivalent to using the assessment of the first two pathologists if they agree and using the third pathologist for resolution if they disagree. The advantage of creating data records in this manner, resulting in 5,145,480 triple reader data records, is that the correct relative weighting is provided to interpretations of cases where the first two pathologist interpretations agree versus where they do not agree, while allowing data to be included for all potential third readers rather than picking one third reader at random from those available when a third reading was required. The 5,145,480 data records resulted from 29 pathologists interpreting the 60 cases in test set A, 27 for test set B, 30 for test set C, and 29 for test set D, yielding a total of 60 × (29 + 28 + 27 + 26 + 25 + 30 + 29 + 28 + 29) = 5,145,480 triple interpretations. Note that for each case the number of simulated triple readings is n × (n − 1) × (n − 2), where n is the number of pathologists who interpreted the case. Figure 1 shows how the triple records were used in conjunction with different criteria for procuring second opinions to arrive at final assessments. We compared the final assessments with...
reference diagnoses to calculate rates of over-interpretation, under-interpretation, and overall misclassification.

In calculating confidence intervals for the over-interpretation, under-interpretation, and overall misclassification rates, we used centiles of the bootstrap distribution of each rate where resampling of pathologists was performed 1000 times. We discarded from the bootstrapped estimates second opinion interpretations that included the same pathologist for second or third interpretations as the first pathologist. P values for the Wald test of a difference in rates between the single pathologist and second opinion strategies were based on the bootstrap standard error of the difference in rates. To calculate \( \kappa \) statistics and rates of agreement between single interpretations we used a simple cross tabulation of all pairwise interpretations of the same cases. The computational burden of analogous calculations for the 5145480 assessments involving second opinions was not tenable so we paired each triple reading of a case with a random permutation of the triple readings of the same case, and, after excluding pairs where the same reader was included in both sets of triples, we calculated agreement statistics. We replicated this procedure 1000 times and report average agreement and \( \kappa \) statistics.

All analyses were conducted using Stata statistical software, version 13.

Results

Table 1 lists the characteristics of the participating pathologists. Forty of 115 pathologists interpreted 10 breast specimens or more weekly and were defined as high volume pathologists. These pathologists were more likely to report spending greater proportions of their clinical time interpreting breast specimens and report that their peers considered them experts in breast pathology.

Among the entire 6900 initial test case interpretations, the pathologists reported that they desired second opinions for 35% (n=2451). Figure 2 shows results by pathologists’ diagnosis of the case. The highest rate of pathologists desiring second opinions was for cases interpreted as atypia (48.4% desired only and 17.7% desired both).

### Table 1 | Characteristics of participating pathologists (n=115) by reported weekly volume of breast biopsy specimens

| Characteristics                        | Total No (%) | No (%), Low volume* | No (%), High volume† | P value‡ |
|----------------------------------------|--------------|---------------------|----------------------|----------|
| **Total sample**                       | 115 (100)    | 75 (100)            | 40 (100)             |          |
| **Age at survey (years):**             |              |                     |                      |          |
| 33-39                                  | 16 (14)      | 11 (15)             | 5 (13)               | 0.98     |
| 40-49                                  | 41 (36)      | 27 (36)             | 14 (35)              |          |
| 50-59                                  | 42 (37)      | 25 (33)             | 17 (43)              |          |
| 60 and above                           | 16 (14)      | 12 (16)             | 4 (10)               |          |
| **Women**                              | 46 (40)      | 28 (37)             | 18 (45)              | 0.42     |
| **Men**                                | 69 (60)      | 47 (63)             | 22 (55)              |          |
| **Experience of breast pathology**     |              |                     |                      |          |
| Fellowship training in breast pathology: | |                     |                      |          |
| No                                     | 109 (95)     | 71 (95)             | 38 (95)              | 0.94     |
| Yes                                    | 6 (5)        | 4 (5)               | 2 (5)                |          |
| **Affiliation with academic medical center:** | |                     |                      |          |
| No                                     | 87 (76)      | 59 (79)             | 28 (70)              | 0.50     |
| Adjunct or affiliated                   | 17 (15)      | 9 (12)              | 8 (20)               |          |
| Primary appointment                    | 11 (10)      | 7 (9)               | 4 (10)               |          |
| **Breast pathology experience (years):** | |                     |                      |          |
| 0-4                                     | 22 (19)      | 17 (23)             | 5 (13)               | 0.45     |
| 5-9                                     | 23 (20)      | 14 (19)             | 9 (23)               |          |
| 10-19                                   | 34 (30)      | 21 (28)             | 13 (33)              |          |
| ≥20                                     | 36 (31)      | 23 (31)             | 13 (33)              |          |
| **Breast specimens as proportion of total clinical case load (% of total clinical work interpreting breast tissue):** | |                     |                      |          |
| 0-9                                     | 59 (51)      | 51 (68)             | 8 (20)               | <0.001   |
| 10-24                                   | 45 (39)      | 23 (31)             | 22 (55)              |          |
| 25-49                                   | 8 (7)        | 1 (1)               | 7 (18)               |          |
| ≥50                                     | 3 (3)        | 0 (0)               | 3 (8)                |          |
| **Considered an expert in breast pathology by colleagues:** | |                     |                      |          |
| No                                      | 90 (78)      | 67 (89)             | 23 (58)              | <0.001   |
| Yes                                     | 25 (22)      | 8 (11)              | 17 (43)              |          |

*Some column percentages do not sum to 100% owing to rounding.
†<10 breast biopsy specimens weekly.
‡≥10 breast biopsy specimens weekly.
Based on Wilcoxon rank sum test for difference in age, experience of breast pathology, and composition of breast specimen of total caseload between low volume and high volume caseload groups. Otherwise P values correspond to a Pearson \( \chi^2 \) test for difference between caseload groups.
When second opinions were desired for specific cases, pathologists noted that 71% (1731/2451) would not be required by laboratory policies in their own clinical practices.

Table 2 shows the rates of agreement with the reference diagnosis after single interpretations and under different strategies for obtaining a second opinion based on characteristics of the initial interpretation. Table 2 also shows for each strategy, the percentage of cases requiring a second pathologist and requiring a third pathologist for resolution of differences between the first two pathologists. The highest misclassification rate within diagnostic categories after single interpretation was for cases of atypia (52.2%), followed by DCIS (15.9%), benign without atypia (12.9%), and invasive carcinoma (3.9%).

The overall misclassification rate for a single interpretation (24.7%, 95% confidence interval 23.6% to 25.8%) was used to compare performance of the different second opinion strategies. Among the strategies described in table 2, the lowest overall misclassification rate resulted when second opinions were obtained for all cases. In this strategy the rates for over-interpretation decreased from 9.9% to 6.0%, for under-interpretation from 14.8% to 12.1%, and for overall misclassification from 24.7% to 18.1%. The percentage of assessments requiring a third opinion for resolution of the diagnosis ranged from 3.7% for invasive carcinoma to 55.9% for atypia. The fraction of assessments where all three readers disagreed was small: 5.1% overall, 2.0% for cases in the benign without atypia reference category, 12.0% for cases of atypia, 3.1% for DCIS, and 0.1% for invasive carcinoma.

Overall misclassification rates compared with the expert consensus reference diagnosis after implementing all of the remaining strategies (table 2) ranged from 19.2% (where second opinion was obtained when desired or required by policy) to 23.9% (where a second opinion was obtained only for cases considered invasive). The only strategy that did not show a statistically significant improvement relative to single reading was obtaining a second opinion exclusively for cases with initial interpretations of invasive breast carcinoma; the overall misclassification rate was reduced from 24.7% on initial interpretation to 23.9% (22.1% to 25.7%, P=0.25). In that scenario, only 10.4% of interpretations required a second opinion, and few (0.4-2.6%) were from reference diagnostic categories other than invasive carcinoma.

Misclassification rates for single readings were lower for cases that were not classified as borderline, difficult, or needing a second opinion (fig 3); however, the misclassification rates for these cases were also reduced when a second opinion was obtained, albeit the improvement was more noticeable for cases that were classified as borderline, difficult, or needing a second opinion.

The second opinion strategies in table 3 were evaluated separately for initial pathologists with a low weekly breast pathology volume and for initial pathologists with a high weekly volume. The overall misclassification rate for single interpretations by pathologists with low weekly volumes was 26.4% compared with 21.5% for those by pathologists with high weekly volumes. The second opinion strategies all showed statistically significant reductions in misclassification rates (P<0.001) compared with single interpretations, with improvement noted not only when the initial pathologist had a low weekly case volume but also when the initial pathologist had a high weekly case volume. The lowest overall misclassification rate, 14.3%, was noted when both first and second opinions were obtained from pathologists with high volumes. The greatest reduction in overall misclassification rate was when the first pathologist was from the low volume group and the second and, if needed, third, pathologist was from the high volume group.

As a secondary analysis we evaluated the agreement rate between assessments of the same case by single readers and assessments of the same case that incorporated second opinions. The average between pathologist pairwise agreement rate for single interpretations of the same case was 70.4%, whereas the corresponding agreement rate for interpretations that included second opinions was higher, at 79.3%. The corresponding k statistics were 0.58 and 0.71.

Discussion
We examined the effects of 12 different strategies for obtaining second opinions by pathologists for breast biopsy specimens. The results support the common belief among clinicians that second opinions should be sought ideally from those with greater clinical experience and especially when the primary reviewing pathologist is uncertain. All strategies showed statistically significant improvements in accuracy except when only obtaining second opinions for cases with...
Table 2 | Over-interpretation, under-interpretation, and misclassification rates of single interpretation compared with reference consensus standard and nine second opinion strategies

| Strategies | Rate (%) | Reference consensus diagnosis | Benign | Atypia | DCIS | Invasive | Overall (95% CI) | P value* |
|------------|---------|-------------------------------|--------|--------|------|---------|----------------|---------|
| **Single interpretation and second opinion applied to all cases** | | | | | | | | |
| Single interpretation | | | | | | | | |
| % requiring 2nd opinion | 0.0 0.0 0.0 0.0 | Benign | 12.9 | 17.4 2.6 | 9.9 (9.0 to 10.8) | 14.8 (13.8 to 15.9) | 24.7 (23.6 to 25.8) | n/a |
| Over-interpretation | – 34.7 13.3 3.9 | Atypia | – | 52.2 15.9 3.9 | 12.9 (11.8 to 15.9) | | | |
| Under-interpretation | 12.9 | DCIS | 27.6 | 216 3.7 | 29.6 | | | |
| Misclassification | 12.9 | Invasive | 29.9 9.3 3.5 | 121 (10.0 to 14.3) | 181 (16.1 to 20.0) | 1.0 0.4 2.6 96.1 | 10.4 |
| **1. Second opinion with resolution applied to all cases** | | | | | | | | |
| % requiring 2nd opinion | 100.0 100.0 100.0 100.0 | Benign | 100.0 | 100.0 100.0 100.0 | 100.0 | | | |
| Overall (95% CI) | – | Atypia | – | 25.7 | 5.0 (3.9 to 6.3) | 14.6 (14.4 to 18.4) | P < 0.001 | |
| P value* | – | DCIS | – | 3.7 | 1.3 (1.1 to 1.5) | 1.3 (1.1 to 1.5) | P < 0.001 | |
| Misclassification | 6.0 | Invasive | 51.9 12.1 3.9 | 21.4 (19.5 to 23.2) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P = 0.25 |
| **2. Second opinion only for initial interpretations considered atypia or DCIS or invasive:** | | | | | | | | |
| % requiring 2nd opinion | 12.9 | Benign | 65.3 | 93.7 99.6 | 61.5 | | | |
| Overall (95% CI) | – | Atypia | – | 26.4 | 4.9 (3.7 to 6.2) | 2.9 | 13.1 12.2 3.3 | 8.8 |
| P value* | – | DCIS | – | 3.5 | 1.4 (1.2 to 1.5) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| Misclassification | 6.0 | Invasive | 51.9 12.1 3.9 | 21.4 (19.5 to 23.2) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| **3. Second opinion only for initial interpretations considered DCIS or invasive:** | | | | | | | | |
| % requiring 2nd opinion | 3.2 | Benign | 17.4 | 86.7 99.6 | 42.1 | | | |
| Overall (95% CI) | – | Atypia | – | 26.4 | 4.9 (3.7 to 6.2) | 2.9 | 13.1 12.2 3.3 | 8.8 |
| P value* | – | DCIS | – | 3.5 | 1.4 (1.2 to 1.5) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| Misclassification | 6.0 | Invasive | 51.9 12.1 3.9 | 21.4 (19.5 to 23.2) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| **4. Second opinion only for initial interpretations considered invasive:** | | | | | | | | |
| % requiring 2nd opinion | 1.0 | Benign | 0.4 | 2.6 96.1 | 10.4 | | | |
| Overall (95% CI) | – | Atypia | – | 26.4 | 4.9 (3.7 to 6.2) | 2.9 | 13.1 12.2 3.3 | 8.8 |
| P value* | – | DCIS | – | 3.5 | 1.4 (1.2 to 1.5) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| Misclassification | 6.0 | Invasive | 51.9 12.1 3.9 | 21.4 (19.5 to 23.2) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| **5. Second opinion only for cases considered borderline or difficult:** | | | | | | | | |
| % requiring 2nd opinion | 19.0 | Benign | 45.3 | 21.4 3.5 | 26.1 | | | |
| Overall (95% CI) | – | Atypia | – | 26.4 | 4.9 (3.7 to 6.2) | 2.9 | 13.1 12.2 3.3 | 8.8 |
| P value* | – | DCIS | – | 3.5 | 1.4 (1.2 to 1.5) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| Misclassification | 9.2 | Invasive | 48.1 11.7 3.3 | 21.1 (19.3 to 22.8) | 14.9 40.4 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| **6. Second opinion only for cases when desired or required by policy or both:** | | | | | | | | |
| % requiring 2nd opinion | 26.7 | Benign | 55.9 | 30.5 15.5 | 35.5 | | | |
| Overall (95% CI) | – | Atypia | – | 26.4 | 4.9 (3.7 to 6.2) | 2.9 | 13.1 12.2 3.3 | 8.8 |
| P value* | – | DCIS | – | 3.5 | 1.4 (1.2 to 1.5) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| Misclassification | 9.2 | Invasive | 47.7 11.4 3.6 | 20.9 (18.9 to 22.8) | 14.9 40.4 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| **8. Second opinion only when required by policy:** | | | | | | | | |
| % requiring 2nd opinion | 14.8 | Benign | 45.3 | 18.0 3.0 | 23.8 | | | |
| Overall (95% CI) | – | Atypia | – | 26.4 | 4.9 (3.7 to 6.2) | 2.9 | 13.1 12.2 3.3 | 8.8 |
| P value* | – | DCIS | – | 3.5 | 1.4 (1.2 to 1.5) | 11.3 43.7 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |
| Misclassification | 8.1 | Invasive | 44.3 10.3 3.7 | 19.2 (17.3 to 21.0) | 14.9 40.4 15.8 3.9 | 217 (19.8 to 23.5) | P < 0.001 |

*Based on Wald test for difference in overall misclassification rates between second opinion strategy and single pathologist interpretation. Test statistic uses bootstrap standard error of difference in rates.
Table 3 | Over-interpretation, under-interpretation, and misclassification rates associated with three second opinion strategies based on case volume of interpreting pathologist

| Strategy | Single interpretation rate (%) | Reference consensus diagnosis | Over-interpretation | Under-interpretation | Misclassification | Second opinion strategy (%) | Reference consensus diagnosis |
|----------|--------------------------------|------------------------------|---------------------|---------------------|------------------|-----------------------------|------------------------------|
|          | Benign                        | Atypia                       | DCIS                | Invasive            | Overall (95% CI)  | Benign                      | Atypia                       |
|          | Low volume pathologist        |                              |                     |                     |                  | 10.5 (9.4 to 11.7)        | 8.4                          | 11.2                          |
|          | Under-interpretation          | –                            |                     |                     |                  | 15.9 (14.5 to 17.4)       | –                            | 30.0                          |
|          | Misclassification             | 14.4                         | 54.8                | 172                 | 4.4              | 26.4 (25.1 to 27.8)        | 8.4                          | 41.2                          |
|          |                               |                              |                     |                     |                  |                            |                              | 10.1                          |
|          | Over-interpretation           | 14.4                         | 179                 | 2.8                 | –                | 10.5 (9.4 to 11.7)        | 8.4                          | 11.2                          |
|          | Under-interpretation          | –                            |                     |                     |                  | 15.9 (14.5 to 17.4)       | –                            | 30.0                          |
|          | Misclassification             | 14.4                         | 54.8                | 172                 | 4.4              | 26.4 (25.1 to 27.8)        | 8.4                          | 41.2                          |
|          |                               |                              |                     |                     |                  |                            |                              | 10.1                          |
|          | High volume pathologist       |                              |                     |                     |                  | 10.5 (9.4 to 11.7)        | 8.4                          | 11.2                          |
|          | Under-interpretation          | –                            |                     |                     |                  | 15.9 (14.5 to 17.4)       | –                            | 30.0                          |
|          | Misclassification             | 14.4                         | 54.8                | 172                 | 4.4              | 26.4 (25.1 to 27.8)        | 8.4                          | 41.2                          |
|          |                               |                              |                     |                     |                  |                            |                              | 10.1                          |

*Based on Wald test for difference in overall misclassification rates between second opinion strategy and single pathologist interpretation. Test statistic uses bootstrap standard error of difference in rates.

High volume pathologists defined as those who report interpreting an average of 10 or more breast biopsy specimens per week. A lower volume pathologist reports 9 or fewer breast biopsy specimens per week. Study sample comprised 75 lower volume pathologists and 40 high volume pathologists.
initial interpretations of invasive breast cancer. Improvements varied according to diagnostic attributes of the cases and the pathologists’ clinical experience. None of the strategies completely eliminated diagnostic variability, especially for cases of breast atypia, suggesting that approaches beyond obtaining a second opinion should be investigated for these challenging cases.

Most US pathology laboratories have policies requiring second opinions for cases of invasive breast carcinoma, yet such diagnoses already have high diagnostic agreement among pathologists. The addition of a second opinion strategy only for cases of invasive breast cancer provided no statistically significant improvement; however, this finding does not indicate there is no clinical value in assuring the highest level of accuracy for invasive carcinoma considering the risks and benefits of treatment. Larger improvements were observed when second opinion strategies included cases with initial diagnoses of atypia and ductal carcinoma in situ (DCIS), two diagnostic categories less often included in laboratory polices mandating second opinions. However, even after applying an array of strategies for obtaining second opinions, the misclassification rates for atypia and DCIS remained high. In actual clinical practice, obtaining second opinions in such diagnostically complex areas might promote, over time, consensus within practices by highlighting diagnostic areas requiring education or expert consultation.

Importantly, a small, proportional reduction in the over-interpretation of breast biopsy specimens may have a large absolute effect at a population level. Obtaining second opinions for all cases with an initial diagnosis of atypia, DCIS, or invasive breast cancer substantially reduced over-interpretation of benign cases without atypia; we observed a reduction in over-interpretation of these cases, from 12.9% with single interpretation to 6.0% with second opinion.

**Strengths and limitations of this study**

This study has potential limitations. The pathologists’ interpretations were independent and only involved a single slide for each case, yet in clinical practice many slides might be reviewed for each case and a second pathologist might be informed of the initial interpretation. It would be impossible and infeasible to design a study of second opinion strategies where full clinical case material for 60 breast biopsy specimens was covertly inserted into the day-to-day practice of more than 100 pathologists from diverse clinical practices. Knowledge of the initial pathologist’s interpretation may influence additional opinions, and this should be studied. Interestingly, about half the pathologists reported that when seeking a second opinion they typically blind the second reviewer to their initial diagnosis.14

We weighted the study cases to include more atypia, DCIS, and proliferative lesions, such as usual hyperplasia, than are typically observed in clinical practice, resulting in higher overall misclassification rates than observed in clinical practice. While the overall misclassification rate is useful for comparing different strategies using second opinions, the results within individual diagnostic categories are more relevant to clinical practice given the weighting of the test cases.25 In addition, future studies should consider differentiating DCIS grade and microinvasion.

It has been suggested that the clinical course of the disease is the ideal means for assessment of the accuracy of a pathology diagnosis.26 However, the clinical course (natural history) is altered by diagnostic excision, clinical treatment, and heightened surveillance after breast biopsy; thus we defined our reference standard as the consensus diagnosis of three experienced pathologists in breast histopathology, a standard acceptable to most women undergoing biopsy and their clinicians. We selected the reference standard as defined by the expert consensus panel after comparison with a reference standard that included the majority opinion of the participants.6 27

The strengths of this study include the large number of participating pathologists (n=115), each interpreting 60 cases from the full range of diagnostic categories, providing 5145480 group level interpretations. We also assessed the impact of 12 different strategies for obtaining second opinions, including obtaining a second opinion on all cases and for predefined subsets based on the initial interpretation. In typical clinical practice, providers often identify challenging cases and then solicit second opinions from the most knowledgeable pathologist (that is, local expert). We simulated this by having pathologists with a high clinical volume provide second opinions in some strategies. For many study cases, participants indicated a desire for a second opinion prior to finalizing their diagnosis; thus, pathologists are likely already obtaining second opinions for cases they encounter in clinical practice, highlighting the relevance of our data.

**Comparison with other studies**

We suspect that patient outcomes will be improved by second opinions in clinical practice, but this was not evaluated. Previous studies have noted consistent rates of discrepant diagnoses uncovered by second opinion within surgical pathology in general,28 and within breast pathology specifically, with second reviews reported to identify clinically significant discrepancies in more than 10% of breast biopsy cases.15 20

**Clinical and policy implications**

Providing second opinions for all breast biopsy specimens or requiring that the interpreting pathologists must be experienced high volume clinicians may be unfeasible given the estimated millions of breast biopsies carried out each year.29 30 Our analysis therefore presents strategies that may be more realistic for clinical practice. Possible barriers to the adoption of second opinion strategies in clinical practice include
workload constraints, uncertainty about the impact of second opinions on clinical outcomes, lack of readily available colleagues with expertise in breast disease, limited reimbursement by third party payers, and concerns about treatment delay. Conversely, the availability of digital whole slide imaging may speed second opinions in the future through telepathology. Though adopting a routine second opinion strategy for some or all breast biopsy specimens may seem daunting, the high rates of disagreement among practicing pathologists on some breast biopsy diagnostic categories is of concern. The financial burden of this variability in pathology diagnoses may be substantial. This includes unnecessary or incorrect treatment, lost income, morbidity, and death.

Conclusion Breast biopsy specimens are challenging to interpret, and many pathologists seek second opinions in clinical practice through informal routes. It might be time for clinical support systems and payment structures to align with and better support clinicians in their current practice. In this study we observed reductions in both over-interpretation and under-interpretation of breast pathology when systematically including second opinion strategies, and we noted that pathologists desire second opinions in a substantial proportion of breast biopsy cases. Improvement was observed regardless of whether a second opinion was or was not desired by the initial pathologist, and it was most notable when the initial interpretation was atypia or DCIS. The feasibility and cost of implementing specific second opinion strategies in clinical practice need further consideration.

AUTHOR AFFILIATIONS

1Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA
2The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Norris Cotton Cancer Center, Lebanon, NH, USA
3Department of Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA
4Fred Hutchinson Cancer Research Center, Seattle, WA, USA
5Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
6Providence Cancer Center, Providence Health and Services of Oregon, and Departments of Medical Informatics and Clinical Epidemiology and Medicine, Oregon Health & Science University, Portland, OR, USA
7Department of Family Medicine, University of Vermont, Burlington, VT, USA
8Department of Family Medicine, Oregon Health & Science University, Portland, OR, USA
9Community and Family Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA
10Rosai J. Borderline epithelial lesions of the breast. Am J Surg Pathol 1991;15:209-21. doi:10.1097/00000478-199103000-00001.
11 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;16:1133-43. doi:10.1097/00000478-199212000-00001.

12 Nakhleh RE, Beikeris LG, Souers RJ, Meier FA, Teorek JA. Surgical pathology case reviews before sign-out: a College of American Pathologists Q-Probes study of 45 laboratories. *Arch Pathol Lab Med* 2010;134:760-3.

13 Tomaszewski JE, Bear HD, Connolly JA, et al. Consensus conference on second opinions in diagnostic anatomic pathology: Who, What, and When. *Am J Clin Pathol* 2000;114:329-35.

14 Geller BM, Nelson HD, Carney PA, et al. Second opinion in breast pathology: policy, practice and perception. *J Clin Pathol* 2014;67:955-60. doi:10.1136/jclinpath-2014-202290.

15 Khazai L, Middleton LP, Goktepe N, Liu BT, Sahin AA. Breast pathology second review identifies clinically significant discrepancies in over 10% of patients. *J Surg Oncol* 2015;111:192-7. doi:10.1002/jso.23788.

16 Newman EA, Guest AB, Helvie MA, et al. Changes in surgical management resulting from case review at a breast cancer multidisciplinary tumor board. *Cancer* 2006;107:2346-51. doi:10.1002/cncr.22267.

17 Elmore JG, Nelson HD, Pepe MS, et al. Variability in Pathologists’ Interpretations of Individual Breast Biopsy Slides: A Population Perspective. *Ann Intern Med* 2016;164:64-9-55. doi:10.7326/ M15-0964.

18 McNair EA, Guest AB, Helvie MA, et al. Changes in breast cancer reports after pathology second opinion. *Am J Surg Pathol* 2002;26:982-7. doi:10.1097/00000478-200204000-00021.

19 Silverstein MJ, Recht A, Lagios MD, et al. Special report: Consensus conference III. Image-detected breast cancer: state-of-the-art diagnosis and treatment. *J Am Coll Surg* 2009;209:504-20. doi:10.1016/j.jamcollsurg.2009.07.006.

20 Tsung JS. Institutional pathology consultation. *Am J Surg Pathol* 2004;28:399-402. doi:10.1097/0107010120034014-3792-5.

21 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 Womens Health* 2013;13:3. doi:10.1186/1472-6874-13-3.