We thank the manuscript committee and reviewers for their valuable comments and suggestions to improve our manuscript. We also appreciate your interest in the topic of diagnostic errors and overdiagnosis and strategies to potentially reduce them.

Our responses to the questions and suggestions of the manuscript committee members and the reviewers are detailed below and we have uploaded our revised manuscript. Please feel free to contact me if you have any additional questions.

Sincerely,

Joann G. Elmore, MD, MPH
Professor of Medicine
University of Washington School of Medicine
Harborview Medical Center
Mailbox 359780
325 Ninth Avenue
Seattle, WA 98104
Telephone: (206) 744-3632
Fax: (206) 744-6097
E-mail: jelmore@u.washington.edu

**Report from The BMJ's manuscript committee meeting**

These comments are an attempt to summarise the discussions at the manuscript meeting. They are not an exact transcript.

Present: Wim Weber (chair); Jamie Kirkham (statistician); Elizabeth Loder; Jose Merino; Rubin Minhas; Alison Tonks; Tiago Villanueva; Georg Roeggla

Decision: Request revisions. Statistician will review when it comes back

* Our statistician noted that you have done a lot of comparisons and many strategies are being compared. We found it difficult to untangle who assessed what and it would be nice to have details. We think a diagram/figure would help to untangle what was assessed and by whom. For example, perhaps you can show how you came to the figure of 191,760 pairs and 5,145,480 triple interpretations from 115 pathologists each interpreting 60 cases.

RESPONSE: We agree that a figure would clarify how we evaluated the different strategies and this has been added to the revision as a new Figure 1. We also agree that it is important to explain how we arrived at the large number of potential pathologists available for second opinions, as this was a unique strength of our study design. In the revised text, we added a description about how the > 5 million group assessments were calculated and the new figure as suggested. Thank you.

* More detail is needed on the allocation of the pathologists to the four groups - details on and group characteristics and group differences might be nice to see (especially if not randomized in some way). As part of the analysis, are there any particular characteristics of the pathologists that can predict better outcome (agreement with the consensus)?

RESPONSE: All pathologists were randomly assigned to one of the four test set groups and the order of case presentation and interpretation for each participating pathologist was also randomly assigned. We expanded the methods section in the revision to add information on our allocation of pathologists to the four test set groups and information on the stratification methods used to assure that the four test sets were similar.

* The expert consensus diagnosis (as a proxy to a gold standard) is an interesting one and a plausible substitute in the absence of a true gold standard. Can you make clearer to readers the rationale for using this as the gold standard, and defend it in some way? Is the overall conclusion that group assessment is better?

RESPONSE: We expanded the discussion section in the revision in response to this important comment. From the perspective of our co-authors who are physicians, our reference standard that involves obtaining opinions from three highly experienced breast pathologists, first independently and then after a consensus discussion using a multi-headed microscope, would be considered a very high ideal in clinical practice. In reality, only the most challenging cases are probably reviewed in consensus discussion at a multi-headed microscope, and this occurs only in larger or academic practices. As investigators, we agree that there is no guarantee that this reference standard, even though based on a panel of three world experts, represents biological truth for all cases.

* Were interrater reliability/Agreement statistics considered as a comparison between the group decision vs. individuals? We would like to see these for individual vs group assessments.

RESPONSE: We added between pathologist percent agreement and kappa statistics for individual assessments and for the group assessments to the results section. This complements our main analysis in Table 2 that evaluates and compares correlations with the expert consensus reference diagnosis.

* This paper will seem somewhat abstract to our clinical readers who are not pathologists. There are no women or outcomes, just some doctors with a lot of slides, which will seem quite removed from practice. The experiment you've done might seem contrived in that it's not typical of real-world second opinion situations where the next doctor might have some clinical background or be given details about the thinking of the first doctor. Perhaps you could briefly discuss these matters.

RESPONSE: Thank you for this question. Millions of women have breast biopsies each year, thus our findings are immediately relevant to these women, their families, and their clinicians. As digital whole slide imaging now makes
telepathology increasingly available, developing methods to study the impact of different criteria for eliciting second opinions will be highly applicable to practice at the health system level.

Our study is analogous to radiologists reading imaging studies. The pathologist is a consultant physician in an ancillary profession. The glass slides, which are a manufactured sample of a patient's tissue biopsy, are the surrogate for the patient. The pathology interpretation is the intermediary outcome between the query that generated the biopsy and the treatment or action that will result from the diagnosis.

Finally, the issues of diagnostic errors and overdiagnosis have increasingly been recognized, with the BMJ boldly taking the lead in this area. This is a challenging topic to study scientifically and politically, thus we may have been overly cautious in describing the clinical implications. The revised manuscript was edited to address your excellent suggestions and we have added a new figure outlining the strategies for obtaining second opinions. These changes improve the clarity and relevance of this study for clinical readers.

First, please revise your paper to respond to all of the comments by the reviewers. Their reports are available at the end of this letter, below. Please also respond to the points above from the committee. In your response please provide, point by point, your replies to the comments made by the reviewers and the editors, explaining how you have dealt with them in the paper.

RESPONSE: We have done our best to reply to all of the comments made by the BMJ manuscript committee and by the reviewers below and feel that the revised manuscript is much stronger. Thank you for this thoughtful and transparent review process.

Comments from Reviewers

Reviewer: 1

Comments:

The report by Elmore et al assessing the impact of second opinion strategies for improving breast Pathology reporting is a continuation of the same work carried out by the same authors and published in March 2015 in the journal JAMA (JAMA 2015;313(11):1122–32). In that study they have assessed diagnostic agreement among pathologists reporting breast biopsies and their publication in JAMA has needlessly undermined public confidence in the practicing community of breast pathologists. There are several limitations in the study methodology that create serious concerns about the generalizability of the study results:

- The study environment was artificial and restrictive: rendering diagnoses as a participant of a research study without clinical implications is remarkably different from the performance of a pathologist in actual practice. Only one pre-selected slide was reviewed per case, and cases were enriched for borderline lesions that are known to have the greatest disagreement. Interpretation in clinical practice is based on examination of all slides of a case and with appropriate clinical context. For the assessment of borderline cases, pathologists often use auxiliary techniques including extra- and immunohistochemistry. Cases of this nature are often the subject of discussion at multidisciplinary meetings and consultation with colleagues as part of the routine reporting process rather than a formal second opinion as envisaged from the current article.

RESPONSE: These are excellent points and we agree that diagnoses of breast carcinoma are frequently discussed at multidisciplinary conferences, particularly in large tertiary care centers; however, benign diagnoses such as atypical hyperplasia are less likely to be discussed in these settings. Our results indicate these cases may benefit from second opinions. Our study provides objective data that can help readers in judging the magnitude of impact that may result from specific strategies for overlaying second opinion in clinical practice.

The topic of diagnostic errors of pathologists is a challenging scientific and clinical area to study and there is no ideal study design. It would be impossible and not feasible to insert full clinical case material for 60 breast biopsy cases in a hidden fashion into the day-to-day clinical practice of > 100 practicing pathologists. This is especially true given that the pathologists are from diverse types of clinical practices from across the U.S. The test cases were enriched with more cases of atypia and DCIS than seen in day-to-day practice to strengthen the study, as it increased our statistical precision for comparisons at the level of diagnostic categories. With that said, we address the implications of the study design in the discussion section. Just because a clinical topic is challenging to study does not mean it should be disregarded.

- Important pathological information is missing. The diagnosis and management of atypia varies between core biopsy and excision specimens. Grade of DCIS and the contribution of microinvasion were not considered in the analysis despite being important in this context.

RESPONSE: These are good clinical points. The interpreting pathologists were aware of whether the case was a core biopsy or excision specimen and we suspect that many took this clinical information into consideration when making their independent assessments of the case. As we noted no difference in accuracy based on type of breast biopsy, we did not include a strategy for obtaining second opinions based on biopsy type. If the Editors would like, we would be happy to add this information to the manuscript.

We also believe that grade of DCIS is an important consideration and mention this in the revised manuscript. The differential diagnosis for low grade DCIS includes atypical hyperplasia and the differential for high grade DCIS, especially when there is fibrosis and architectural distortion. One of our cases of DCIS included a small focus of microinvasion on excision biopsy and study participants frequently missed the microinvasion; however, the reference panel did not believe this finding would have changed management of the case. Microinvasion on a core biopsy might influence a clinical decision regarding axillary node evaluation if that were not already being considered based on high-grade DCIS and imaging features. While the complexity of clinical decision-making and the value second opinions might contribute on an individual case basis is difficult to incorporate into a large analysis such as the Breast Pathology Study, we believe we have been able to objectively demonstrate the general value of second opinions in diagnosing atypia and DCIS.

- Discordance in the reporting of atypia may not mean disagreement. In real practice the same diagnosis in different institutions may trigger vastly different actions. Conversely different diagnoses may result in identical treatment. Hence the discordance does not equate to right-wrong diagnosis nor necessarily harm to the patient.
Grey areas exist in all fields of medicine and publication of such an article in BMJ is expected to increase public concerns regarding the performance of breast pathologists in routine practice and undermine their valuable contribution to the management of patients with breast disease. Addressing the limitation of the study in the same manuscript or in separate letters to editors will not eliminate the unnecessarily and unfounded anxiety the publication of the article will create.

**RESPONSE:** These are excellent points and truly demonstrate the dilemma clinicians face in translating objective study outcomes into clinical practice, and the dilemma researchers face in trying to model clinical practice in research studies. The science and art of medicine need to be balanced, and published studies inherently tend to focus on the scientific aspects of an issue.

We agree that pathologists are important and often pivotal contributors to the management of breast disease. Our intent is not to undermine this contribution or the public confidence; instead, our goal is to understand how we might improve the process. The first step is to assess the current state. We hope that readers will integrate both the limitations and the objective findings in applying the data toward improving clinical care for the millions of women having breast biopsies each year. Without periodic assessment, it is difficult to implement systemic improvements. We would also hope that any short-term repercussion of public anxiety would be offset by maintaining long-term focus on quality improvement including a better understanding of how we might tailor management for diagnostically grey areas of breast pathology.

The paper is very much suited for a pathology journal addressing this important topic to the breast pathologists who are the right audience for such articles rather than going to the readers of BMJ.

**RESPONSE:** Our study findings are relevant to the millions of women and their clinicians dealing with the results of breast biopsies each year and to their health systems contemplating strategies to improve care and quality while also considering cost and workforce capacity. Our study design is also innovative and can be used as a model going forward as the field of telemedicine (including telepathology) opens up broader access to second, and third, opinions in medicine.

**Reviewer:** 2
**Comments:**
Diagnostic interpretation of breast biopsy specimens is of great significance for people (women/men) with breast lesions. The breast pathology diagnosis provides the basis for clinical treatment and management decisions; that being the case, diagnostic accuracy is substantially important to ensure the most applicable therapy plan. Misclassification of identified lesions may result to either overtreatment or undertreatment. These issues are particularly important if we consider the widespread mammography screening and the consequently increasing incidence/detection of breast lesions. Millions of breast biopsies are performed annually around the world. The economic, psychological and societal burden of mistreatments could be enormous.

Based on such research both lay people and doctors can be educated in the management of a health crisis, reduce anxiety and decrease the burden of misclassification and mistreatment. A second opinion adds value and it is desirable. Lay people learn to be involved in management of their own health and to contribute effectively with doctors in their health plan. Additionally, they get duly informed of the eligible criteria of research (low vs high volume doctors and/or academic affiliation) in order to make the best health decisions.

Last but not least, this research highlights some system failure. For example, pathology laboratories’ policy to require a second review mainly in case of diagnosed invasive cancer where the diagnostic disagreement is significantly lower than other cases such as atypia and DCIS. Although such reviews are requested by pathologists themselves and they are reported to improve their diagnostic accuracy. Furthermore, the majority of pathologists hasn't completed fellowship training in breast pathology.

**Note:** regarding the strategy that has been implemented in case of total disagreement of diagnostic interpretation among (3) doctors I consider that further estimation should be applied (i.e. arbitrary assigning ambiguous diagnoses in the middle of the diagnostic hierarchy arises ethical considerations).

^1 pages 35,38, figures 1,3 &
^2 page 6
^3 page 37, table 2
^4 page 10 BPATH-Dx

**RESPONSE:** Thank you for your comments and for reviewing our manuscript. We used majority opinion and median opinion as statistical methods to evaluate effect. In clinical practice, when a definitive consensus diagnosis after consultation cannot be achieved, we anticipate that the individual case situation and implications of the varied diagnoses would be considered in any individual patient’s treatment or follow-up plan. This degree of complexity of clinical decision making would be extremely difficult to model in a large study.

We were pleased to see a patient expert as a reviewer at The BMJ, especially given the importance of this topic to patients, their families and their clinicians, in addition to the overall health care system. We have breast cancer survivors among our investigator group and also among our family members. We appreciate your time and encouragement to help improve the quality of care we provide to all patients. Getting a correct diagnosis is an important first step.

**Reviewer:** 3
**Comments:**
Interpretations from 115 pathologists of 240 breast biopsy specimens (atypia, DCIS, and invasive cancer, one slide per case, were used to establish baseline accuracy of single observers’ diagnoses. These were compared to accuracy based on independent interpretations by pairs of pathologists (N=191,760 paired interpretations). The authors then evaluated twelve strategies aimed to reduce misclassification rate, with the "true" classification based on expert review of 3 expert breast pathologists. This is an informative and detailed study of a very clinically relevant problem encountered frequently in daily practice.

The strategy of creating pairs and triplets for the analysis is clever and statistically efficient. However, the title is misleading in that the study actually compares the misclassification rate if the first and second opinions agree and the accuracy of the third second opinions disagree.

**RESPONSE:** Thank you for your kind comments about our study design and the importance of this clinical topic. We agree the title should be improved. The new title we propose is: "Evaluation of Criteria for Obtaining Second Opinions to
Improve Breast Histopathology Interpretation.” If the second reader agrees with the first, then the final assessment is their common assessment. If the first two readers do not agree, then a third is brought in and the majority assessment is the final assessment (or the middle assessment if all three disagree). We have added a new Figure 1 to describe our methods based on feedback from the Editors. It is not possible to fully describe this definition in the title and we feel it is adequate to simply call it an “Evaluation of Criteria” for obtaining second opinions. We are open to any further suggestions for the title.

Were there cases where there were 3 separate and different opinions? If so, how were these cases handled—were they excluded from the analysis? 
RESPONSE: Yes, three different diagnoses were occasionally provided on the same case. This occurred 5% of the time among all of the possible triple combinations (there were over 5 million possible triple readings). When all three pathologists disagreed, we selected the middle diagnosis as the final. These rates differed by reference diagnostic categories: benign without atypia 2.0%; atypia 12.0%; DCIS 3.1%; and invasive breast cancer 0.1%.

Most of the misclassification appears to arise in the diagnoses of atypia and DCIS. How often did the 3 expert opinions differ in their adjudication of atypia vs DCIS? The most clinically significant misclassification lies in upstaging of atypia to DCIS or downstaging of DCIS to atypia. Can any conclusions be reached about those cases which were most frequently misclassified?
RESPONSE: These are interesting questions. Misclassification was noted among all of the cases, without a subset of specific cases standing out as having the most misclassification. We also refer this reviewer to Table 1 and the subsequent Letter to the Editor for our 2015 JAMA publication regarding data on the three expert reference pathologists. Biopsy cases of atypia are the most challenging for all pathologists, including the experienced breast pathologists.

As the majority of breast biopsies each year in the general population are diagnosed as benign without atypia, over-interpreting benign tissue as atypia is perhaps a bigger clinical issue because this could result in overall increased surveillance and potential overdiagnosis and overtreatment for a large number of women. We also agree that a very important misclassification is “upstaging” (over-interpretation) of atypia to DCIS given the treatments that women with DCIS receive.

“Missing” DCIS by calling it atypia is probably not as clinically significant because the under calls would most likely represent low grade DCIS, a diagnosis that does not necessarily need immediate intervention. If this progresses, it would likely be detected on future screening once it is more easily recognized.

Were there any “outliers” who consistently achieved highest concordance with expert review? If so, can any conclusions be reached about this group?
RESPONSE: Pathologists from academic centers, those with high interpretive volumes, and those who work in larger sized clinical practices were more likely to have higher concordance with the expert consensus reference diagnosis. We suspect that pathologists who work in larger practices (defined as having >= 10 other breast pathologists in their group) have higher concordance with the reference standard because they are able to easily ask for and receive second opinions on cases.

Reviewer: 4
General comments:
This manuscript by Elmore J et al presents different strategies for adding independent second opinions for improving breast biopsy interpretation amongst participating pathologists. In their earlier study, they assessed diagnostic concordance rates amongst the pathologists with an expert consensus derived reference diagnoses. They agreed on lowest concordance rate for cases involving atypia and DCIS (Ductal carcinoma in situ) and highest concordance rates for cases with invasive carcinoma. In this study, the authors present twelve different strategies for obtaining second opinions in breast biopsy interpretation and find that all of the strategies significantly improve the overall misclassification rate except for cases with invasive cancer. They also note that diagnostic variability is not completely eliminated especially for cases reported as atypia. Although many studies exist in literature that have looked at the role of second opinions in pathology case interpretations, this study presents different strategies for obtaining second opinions which has not been looked at in prior studies. This study is also clinically relevant as many pathologists had also indicated a desire for second opinion especially in cases that are not required to do so by policy, recognizing the need for better review strategies in interpreting diagnostically challenging cases.

1. Research question:
The hypothesis in this study has been well presented with clear mention of the research question i.e. an analysis of various second opinion strategies and whether they affect breast pathology interpretation.

2. Originality:
Many studies have looked at usefulness of obtaining second opinions for breast pathology interpretation, with some of these uncovering clinically significant discrepancies in reporting breast biopsy cases. However, the authors are correct in stating that none of them have compared or at least identified different systematic strategies for obtaining second opinions in breast pathology.

3. Overall Study design:
The study has a well presented and a strong research design.

4. Study sample/ participants:
I feel that this is adequate. The inclusion and exclusion criteria are clearly described.

5. Adherence to standards:
Adequate, the authors have also included a completed STARD checklist. The manuscript is also free of typographic errors to the best of my review with up-to-date and matching references.

My overall impression is that this is a clearly executed, well-thought of and presented study on an important topic of concern in the medical field. It is suitable for the general reader audience of the BMJ and definitely merits consideration for publication.

Further comments follow.

Specific comments (strengths, suggestions for improvement and/or clarifications):

1. Abstract:
a. Page 3 (lines 33-38), the authors gives us a description of a total of 240 cases interpreted. Perhaps, it would be better to include the numerical 240 in the first line of the methods section reading as “Interpretations from 115 pathologists of 240 breast biopsy specimens” so that the reader can better understand succeeding lines.

b. Remainder of the abstract including the conclusion is well written and to the point.

RESPONSE: While we did indeed have 115 pathologists, they did not all interpret 240 test cases. Each pathologist was randomized to a test set and interpreted 60 total test cases. Because of this point, we listed the total number of test cases in a different sentence in the abstract.

2. Introduction:

a. While the authors are correct in saying that there are no studies comparing different systematic approaches for error reduction in breast pathology interpretation, utilization and comparison of such strategies has been described in other areas of pathology. Perhaps, the authors may want to consider mentioning some of the other work regarding second opinions in other areas of pathology that have been done. For e.g. in the article by Laban et al, they have proposed the use of algorithms for dealing with diagnostic difficulties in cutaneous lymphoproliferative disorders utilizing regional network resources.

b. It is also notable that the best methods for obtaining 2nd opinions in pathology are not known and I agree with the authors in stating this. Perhaps, with more studies comparing such different strategies, institutions can come up with revised protocols for obtaining second reviews especially in diagnostically complex areas.

c. Page 14 (lines 13-18), the authors state that the lowest overall misclassification rate was noted when both first interpretation and second opinion were obtained from high volume pathologists. While this is true, the greatest reduction however seems to be when the first interpretation was obtained from low volume pathologists and second and third opinions from high volume pathologists. The authors may want to clarify this.

RESPONSE: This point has been mentioned in the revised manuscript.

b. From each of the strategies discussed, it also seems clear that cases of atypia and DCIS tend to be under-interpreted when considered as an interpretation rate % of the reference consensus diagnosis. Perhaps, the authors would like to comment on this observation regarding high under-interpretation rates.

RESPONSE: This is an important point. As most breast biopsies in practice are benign without atypia, a higher total number of breast biopsies are over-interpreted at a population level. We have added a new reference regarding this point in the discussion section (J Elmore et al, Ann Internal Medicine 2016).

c. In their previous investigational study (2), the authors state that “Disagreement with the reference diagnosis was statistically significantly higher among biopsies from women with higher (n = 122) vs lower (n = 118) breast density on prior mammograms (overall concordance rate, 73% [95% CI, 71%-75%] for higher vs 77% [95% CI, 75%-80%] for lower, P < .001). It may be interesting to know if predetermined variables such as these may affect formulation of strategies for obtaining second opinions.

RESPONSE: The topic of breast density is receiving intense interest, especially as it relates to breast cancer screening practice in the U.S. While we found breast density to be associated with pathologists’ diagnostic accuracy, the magnitude was small. We evaluated the strategy of obtaining second opinions when women have high-density breast tissue, as suggested. We note that our design fixed the number of cases with high-density breast tissue at about 50%. We calculated the overall misclassification rate when second opinions were procured for all cases with dense breast tissue (20.9%). Not surprisingly, since this rule sends about 50% of cases for second reads, the misclassification rate (20.9%) is between the misclassification rates when no cases get second opinions (24.7%) and when all cases get second opinions (18.1%). Most revealing is the fact that using low breast density as a trigger yields about the same misclassification rate (21.9%) as when using high breast density as a trigger (20.9%). Thus, breast density in and of itself does not seem to be helpful in dictating which cases should go for a second opinion.

We suspect that the association of pathologists’ accuracy with mammographic density is more likely associated with breast epithelial proliferation, a factor that has not been sufficiently studied with respect to causal associations since the major component of mammographic density is fibrous tissue.

d. Page 15 (lines 20-25), the authors state “In actual clinical practice, obtaining second opinions in such diagnostically complex areas may, over time, promote intra-practice consensus by highlighting diagnostically areas requiring education or expert consultation.” Some studies such as the one mentioned below have been conducted in Europe assessing the impact of forming regional networks to assist in diagnosing difficult lesions of the breast, with the goal of achieving standardization across pathological processes.

RESPONSE: While we did not add this citation (the full paper is in French), but agree that the network approach supporting
clinical care for these challenging cases may improve practice. We suspect that more health systems will move to this approach to clinical care in the future, thus we evaluated multiple possible criteria and strategies in our study. We have noted that pathologists who work in practices with \(\geq 10\) other pathologists who also interpret breast biopsy specimens have higher agreement with the reference standard. We suspect that this is due to their ability to discuss cases with colleagues in their day-to-day clinical practices.

e. The conclusion section of the manuscript is also well written and adequately summarizes the study with identification of areas for conducting further research.

5. References
The references are generally well written and comply with the journal’s standards.
Page 21 (line 8) Should read as JAMA 2015;313(11):1109-10
RESPONSE: Thank you.

6. Summary/ Key message
Overall well conveyed.