Author's response to reviews

Title: Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: Impact on relative risk of breast cancer

Authors:

Rikke R Winkel (rikkerass@dadlnet.dk)
My von Euler-Chelpin (myeu@sund.ku.dk)
Mads Nielsen (madsn@diku.dk)
Pengfei Diao (diao@diku.dk)
Michael B Nielsen (mbn@dadlnet.dk)
Wei Y Uldall (wei.uldall@regionh.dk)
Ilse Vejborg (ilse.vejborg@regionh.dk)

Version: 2
Date: 25 January 2015

Author's response to reviews: see over
Response to reviewer comments regarding:

Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer

(Title has been changed)

Reviewer 1: Alberto Tagliafico

Reviewer's report:

Major Compulsory Revisions

1. Page 6 line 3-5: not possible to cite only Cumuls. Cite also Volpara, Quantra, MedDensity and add references and introduce the concept of breast density and digital breast tomosynthesis:


3: Tagliafico A, Tagliafico G, Astengo D, Airaldi S, Calabrese M, Houssami N.


Answer: We appreciate this comment and thank the reviewer for pointing out relevant references. Numerous newer semi or fully-automated computerized methodologies exist (measuring area or volume-based measures; percentage and total glandular density; measures of density or texture). Computer-aided techniques have been shown to have higher reproducibility among readers and automated methods have the advantage of not being subjective.

In the background page 6 line 11-12 we point out that the present study is part of an ongoing research project validating a new automated computerised density score and a new automated texture score (on digitized analogue mammograms). The main purpose of our present study was to focus on and compare the methodologies involving a subjective component, in order to investigate the impact of subjectivity on relative risk estimates of breast cancer in terms of ORs. We believe it is beyond the scope of this article to, in great detail, discuss different automated techniques and other modalities. But we thank the reviewer for pointing out relevant references and agree that the newer automated (including volumetric) methods—and modalities—should be cited (and discussed in brief).
We have already cited Volpara and have now added references to Quantra and MedDensity as well as briefly noted other upcoming modalities for density assessment.

**Action:** **Background: Line 1-9, page 6:** The last part of the section has been re-written and now reads: “Hence, partially and fully-automated computerized techniques are an area of active research. Several computer-aided techniques exist where the interactive area-based commercialized Cumulus software is most commonly used [29]. However, subjectivity is still not completely eliminated by the partially-automated techniques. Thus, research has in recent years focused more intensively on a fully automated objective assessment of breast density, including volumetric measures, in line with breast imaging moving from analogue to digital mammography [30–33]. In addition, density assessment carried out using other imaging modalities as digital breast tomosynthesis (DBT) or MRI are also being investigated [34, 35].”

2. **Methods**

It is not reporter if subjective density and parenchymal pattern classifications were investigated with randomization (a) of the images and it is not clear how the readers recuded the bias of assessing the different patterns of mammogramms without knowledge of the first evaluation (b). This issus increases artificially the agreement and reduces a lot the importance of the study. Please clarify VERY accurately the study design regarding image evaluation .

The Authors say that: The three classifications were also completed at different stages in order to prevent the reader from being influenced by the assessment by other classifications. This is not sufficient to explain the study design and to understand how much the bias has been reduced. Indeed the results seems too optimistic and not clinically likey to be reproduct.
**Answer:** a) Density and parenchymal classifications were done on negative screening mammograms (no cancer detection) which were randomized in regards to later development of cancer among cases. We agree with the reviewer that it should be specified that the images were randomized in respect to case-control status.

b) The readers evaluated the images in a MatLab scoring-database which was designed specifically for the reading purpose. Using this program the reader had to choose density scoring by BIRADS, Tabár or PMD. When one method was completed a new method was chosen. Thus, the readers were blinded from previous scorings by the different methodologies. Scorings were done over a period of six months, with an average of one month between each method. We agree with the reviewer that this can be specified more clearly.

**Action: Methods:** a) Line 12-14 page 8: The sentence now reads: “The digitised mammograms were randomized according to case/control-status and reviewed independently by two medical doctors: a senior radiologist specialized in breast-imaging and mammography screening (Reader 1) and a resident in radiology (Reader 2).”

b) Methods: Line 12-18 page 10: The section now reads: “Readings of one breast-side of all the women were completed before scoring the opposite breasts (never evaluating a woman’s right and left breast together). Accordingly, the right and the left breasts were scored separately and can thus be considered independent measurements. Readings by the three different methodologies were completed separately at different times over a period of six months in a MatLab scoring-database. In order to further reduce artificial agreement between the methods, the readers were blinded from evaluations by the other classifications.”

3. Please add intra-observer agreement at least in a percentage of cases.
Answer: We agree with reviewer that this could also be of interest. Unfortunately, it is not possible to do an *intra*-observer study regarding the BIRADS and Tabár classifications. The three methodologies investigated in this study are also being compared with two fully automated techniques in an ongoing study (we kindly refer to our answer on comment 1). For this purpose we have used consensus scores on BIRADS and Tabár. The consensus readings have lead to one reader “tuning in” on the second reader. This would lead to bias if an *intra*-observer study were to be performed. Since it is also normally assumed that *inter*-observer variation is bounded by *intra*-observer variation, *intra*-observer agreement would likely be higher. This is why we haven’t conducted an analysis on *intra*-observer reproducibility.

4. Discussion


Answer: We kindly refer to our answer on comment 1. We have added references and elaborated briefly on the subject in the Introduction (see above) and Conclusion.

Action: Conclusion: Page 23 line 17 – page 24 line 3: The section has been re-written: “Thus, an automated, objective and reproducible method to estimate density or texture (or both) from the
mammogram are needed to fully overcome the impact of subjectivity. Our study is based on analogue images. However, many breast imaging units have in recent years switched to digital mammography. This has encouraged the development and improvement of fully automated techniques, which has been shown to be valid alternatives on digital mammography [32, 33]. In addition, the applicability of other imaging modalities for density assessment is being investigated including DBT and MRI [34, 35, 51]. The numerous methodologies existing today may capture different aspects of density, and it remains unresolved which particular methods to use. This will necessarily depend on the aim (research/clinic/tailored screening). However, it is evident that different methods are not interchangeable.

In conclusion, our study confirms that improvement of fully automated methods should be continued to overcome subjectivity (as well as time consumption) in measuring density for research and clinical risk assessment.”

5. Too many results are repeated in the discussion, shorten a little bit.

Answer: We agree with the reviewer that some results do not need to be repeated in Discussion. We have now shortened the Discussion some and removed repeated values. We have also shortened the Results a bit where numbers are clearly presented in tables.

Action: The Discussion has been shortened and repeated values have been removed—please see revised manuscript.

6. References

not compete, see above
**Answer:** We appreciate reviewer’s input and agree that more methodologies and imaging modalities could be referred to and briefly discussed, even though, this paper specifically focus on subjective methods and the impact of subjectivity on relative risk estimates. We kindly refer to our action on reviewer comment 1 and 4.

**Action:** Background: Page 6 line 7 and 9: We have added the following references:

32. Tagliafico A, Tagliafico G, Tosto S, Chiesa F, Martinoli C, Derchi LE, Calabrese M: 


**Conclusion:** Page 23 line 21 and 23: We have added the following references:
32. Tagliafico A, Tagliafico G, Tosto S, Chiesa F, Martinoli C, Derchi LE, Calabrese M: 
Mammographic density estimation: comparison among BI-RADS categories, a semi-automated 


34. Tagliafico A, Tagliafico G, Houssami N: Differences in breast density assessment using 
mammography, tomosynthesis and MRI and their implications for practice. Br J Radiol 2013, 
86:20130528.

35. Tagliafico A, Tagliafico G, Astengo D, Airaldi S, Calabrese M, Houssami N: Comparative 
estimation of percentage breast tissue density for digital mammography, digital breast 

51. Tagliafico A, Tagliafico G, Astengo D, Cavagnetto F, Rosasco R, Rescinito G, Monetti F, 
Calabrese M: Mammographic density estimation: one-to-one comparison of digital 
mammography and digital breast tomosynthesis using fully automated software. Eur Radiol 

7. Table and Figures

use the same numbers of decimals: one, two, three???
**Answer:** We thank for pointing this out and agree that the same number of decimals should be used in the tables and the text when reporting the same results. We also agree that the same number of decimals should be used describing the same values (e.g. p-value)

**Action:** P-values have been made uniform using three decimals. Kappa and ICC values are now reported with two decimals in both tables and text. Percentages are reported with none or one decimal (depending on what the value describes).

**Reviewer 3: steven allen**

**Reviewer's report:**

This is a well written paper which is largely scientifically sound although I have a few but important points to make.

1. In looking at 3 subjective methodologies, this paper is of interest but not scientific importance as newer automated methodologies have been shown to be far superior in reliability of density assessment. This reference is a case control study to this effect "Digital mammographic density and breast cancer risk: a case control study of six alternative density assessment methods." Eng, A. ; Gallant, Z. ; Shepherd, J. ; McCormack, V. ; Li, J. ; Dowsett, M. ; Vinnicombe, S. ; Allen, S. ; Dos-Santos-Silva, I. ; Breast Cancer Res, 2014; 16(5):439

**Answer:** We appreciate the comment from the reviewer and we thank the reviewer for pointing out the reference. In this reference the association between **FFDM**-based density and the risk of breast cancer is investigated regarding six different methods. All methods have overlapping CIs, so it is difficult to conclude that one method is better than the others, even though Volpara has the highest ORs. This is also consistent with the literature on the area in general.
We find our study to be of scientific importance because:

1) We use film-based mammograms in our study. Even though digital mammography is now widespread, film-based mammography still exists worldwide. Moreover, there is important information on density and parenchymal pattern from old analogue mammograms, and we believe this will be of interest for epidemiological long term follow-up studies for many years to come. Density cannot easily be assessed by automated techniques on these film-based images as raw data rarely are available in retrospective studies.

2) To our knowledge our inter-observer study is the first to report inter-observer agreement on the Tabár classification. In general, reproducibility on Wolfe or Tabár patterns is seen as being low by the scientific community. However, in this study we find reproducibility to be comparable with the well tested density methods (without communication with Tabár himself). We find this to be a very important message of this study, particularly due to the upcoming discussion about risk and masking regarding structural density. Our other results verify what has already been shown in the literature, and can be thought of as supportive material: it shows that our study hasn’t any major errors, since we get expected results.

The three methods investigated in this study are being compared with two fully automated techniques in an ongoing study (we kindly refer to comment 1 by reviewer 1). In that study we focus on the difference between assessing density or texture/parenchymal pattern of the breast—validating a new mammographic texture technique. We aim to investigate if the two approaches might complement each other. Moreover, we compare the new mammographic texture technique with the Tabár classification (representing more than just a quantitative density measure - but density structure).
3) Importantly, the subjective BIRADS density classification is still the most used method for density assessment in daily clinical practise and for tailored screening. However, the new BI-RADS density definition (5th edition) is no longer quantitative but qualitative. Therefore we cannot just immediately use volumetric density. ACR has thus made subjective (structural) assessments more important in the future.

4) Finally, we believe the present study also contributes to clarifying what impact the use of these subjective methods can have on breast cancer risk estimation. Our study supports that automated methods are warranted. Please note that the cited reference was not available at the time of submission of this paper. We have now added this reference (ref 33).

Action: We have added some of this discussion under Strengths and Limitations Page 21 line 18 to Page 23 line 2: “...picture of the degree of agreement as discussed in detail by Abdolell et al [49]. Moreover, we find it a strength of this study to have included the qualitative Tabár classification and demonstrated its reproducibility. With ACR’s new definition on the BI-RADS density classification (5th edition) returning from a more quantitative to a qualitative classification, it seems as if the more qualitative classifications also have a role to play in the future.

We recognise our study also has some limitations to be addressed: In this retrospective study on a screening cohort we have not been able to control for other breast cancer risk variables other than age. However, from a clinical point of view the question is what we can do with the information available to us, if we were to do risk-based stratification of screening women. In many screening programmes—like ours—the only information available to us is the woman’s age and her mammogram. Therefore, ORs have not been adjusted for other risk factors such as BMI, history of breast cancer, menopausal status, and other reproductive variables in this study. The ORs should
obviously be interpreted with precaution when compared with other studies, and are in the present study primarily to be compared between readers. BMI is known to be one of the most important confounders; however, the lack of adjustment for BMI has probably led to some underestimation of risk [4, 50]. Moreover, we did not differentiate between interval cancers (defined as cancers diagnosed between two screenings) and screen-detected cancers. We might have included some “excess” cancers which may have been initially un-detected (masked at the negative screening in 2007), leading to an overestimation of risk [4].

In addition, readings were done on analogue digitized mammograms reducing the quality of the images. Mammograms were rather dark and, accordingly, the breast boundary was not easy to delimit and might have influenced PMD estimation. The readers also had to compensate for colouring artefacts (e.g. from the pectoral muscle) when setting the threshold. Accuracy and reliability of methods for density assessment on digital mammograms (including automatic techniques) may be superior. However, important information from film-based mammograms still exists. We believe this will be of interest for epidemiological long follow-up studies for many years to come.

Finally, it would have strengthened our study methodologically to have had more readers. Keeping the above limitations in mind we did find our results to be comparable to others, though.”

2. Another major flaw in this paper is that there has been no assessment made of non breast density breast cancer risk other than age. It is alluded to in the discussion that BMI was not obtainable and clearly this is an important example of this. However there are many other
questionnaire type breast cancer risk variables such as menopausal status, contraceptive use etc. all of which are not controlled for in your study.

With the above in mind I therefore cannot accept in anyway your results/conclusions about breast cancer risk in relation to breast density. I would consider this a major flaw in the study and therefore the only result of interest to me is the comparative analyses between the different methodologies.

**Rebuttal:** We appreciate the comment and we agree that the wording in the manuscript is imprecise. In our study we are not investigating exact risk prediction, but a relative risk of breast cancer for women with certain breast density patterns in a screening population in terms of ORs. The main purpose of our study was the comparison of the two readers (and also to a limited extent the comparison of methods). We intended to compare reproducibility, and secondarily the potential effect on relative risk estimates of breast cancer in terms of ORs adjusted for age between two readers. We have now changed the wording throughout the manuscript from risk prediction to relative risk of breast cancer.

This is a retrospective study and information on other breast cancer risk factors (or confounders) was unfortunately not available to us (since this information is not collected for screening in Denmark, and probably not will be in the near future). From a clinical point of view the question is what we can do with the information available to us, if we were to do risk-based stratification of screening women. In many screening programmes—like ours—the only information available to us is the woman’s age and her mammogram.

Moreover, ongoing research, also using data from stemming from the Copenhagen screening program, has shown that the association between density and breast cancer is robust. The relative risk of breast cancer for women with mixed or dense breast tissue in comparison to fatty tissue
changed from 1.94 (1.47-2.57) to 1.81 (1.34-2.44) when adjusted for education, alcohol use, physical activity, BMI, parity, number of children, age at first birth, and history of benign breast tumor. (Personal communication: Andersen ZJ et al). Regarding Tabár the specific association with PIV was also found by Jakes and colleagues (Jakes RW, Duffy SW, Ng FC, Gao F, Ng EH: Mammographic parenchymal patterns and risk of breast cancer at and after a prevalence screen in Singaporean women. Int J Epidemiol 2000, 29:11–19.). They demonstrated an unadjusted OR of 2.59 (1.31-5.13) when PIV was compared with the combined group of Tabar’s pattern I, II, III and V, which was also seen consistently (and significantly) after adjusting individually for other breast cancer risk variables and confounders. We have now tried to specify this under Methods and in the Discussion under limitations.

**Action:** We have changed the wording throughout the manuscript from risk prediction to relative risk of breast cancer (in terms of ORs).

**Methods Page 12 line 11-12:** The sentence have been re-written and now reads: “Due to the retrospective design of this study, information on body mass index (BMI) and other breast cancer risk variables could not be obtained and controlled for.”

**Discussion Page 22 line 1-9:** We have re-written the following section under Strengths and limitations: “We recognise our study also has some limitations to be addressed: In this retrospective study on a screening cohort we have not been able to control for other breast cancer risk variables other than age. However, from a clinical point of view the question is what we can do with the information available to us, if we were to do risk-based stratification of screening women. In many screening programmes—like ours—the only information available to us is the woman’s
age and her mammogram. Therefore, ORs have not been adjusted for other risk factors such as BMI, history of breast cancer, menopausal status, and other reproductive variables in this study. The ORs should obviously be interpreted with precaution when compared with other studies, and are in the present study primarily to be compared between readers.”

We also kindly refer to our discussion regarding comment 1.

More specific points are as follows:

3. With regards automated analyses there are several papers including the above that cite these techniques and show their accuracy. This should be mentioned more extensively in relation to both the introduction and discussion.

Answer: We thank for this comment and for pointing out the reference. We kindly refer to the first, fourth and sixth comment by reviewer 1. The above paper by Eng et al has now been cited.

Action: We kindly refer to the action corresponding to the first, fourth and sixth comment by reviewer 1, where changes have been made in the Introduction and Conclusion including more references.

4. Your follow-up data concludes at 3yrs which is a short timescale. Why was that used?

Answer: We agree that it would have strengthened the study with a longer follow-up period. The reason why 2007 was chosen is because the acquisition of images became more standardized from the year of 2006. From 2006 it was mandatory to require both MLO and CC views at screening. Before 2004 women with fatty breasts had only MLO views taken except from at their first screen (2004-2006 transitional period).
5. December 2010 was a long time ago. Why have you not published prior to this?

**Answer:** Early reporting of results is of course always preferable. The delay can in part be explained by the necessary procedures in acquisitioning and digitalization of images. Cases and controls were collected in the first half of 2011. It took approximately two years to retrieve the film-based images from the hospitals archive, digitize all the images and prepare the scoring database. Approximately one year were used on density assessment and analysing data. The article was submitted in mid February 2014, and we received our first reviewer comments in end December 2014.

6. Within the discussion you allude to the masking effect (**p17 line 8**). Please explain this further

**Answer:** We value this comment and have now elaborated on this. The section has been re-written.

**Action:** Discussion **p17 line 3-9.** The section now reads: “The differentiation into high/low-risk categories is central as it has been suggested to form the basis of personalized screening with particular attention to the masking effect [6, 23]. Mammographic sensitivity decreases in line with increasing breast density due to superposition of overlapping normal breast tissue and potential breast lesions. This masking effect on two-dimensional images leads to increased risk of interval cancers. Accordingly, women with high density may benefit from supplementary exams with e.g. digital breast tomosynthesis in which the breast is viewed in “slices” or “slabs.””
7. Discussion page 18, line 12. I think it is an assumption to assume Tabar classification familiarity is affected by reader experience especially when you describe that none of the readers had used this technique prior to the study.

**Answer:** We appreciate pointing this out and agree with the reviewer. Tabar classification familiarity is not affected by the reader’s breast imaging experience. However, the interpretation/perception of the classification is most likely dependent on experience. The sentence has been re-written.

**Action:** Discussion p18 line 12-13. The sentence now reads: “Again, this is largely a matter of perception of the mammographic structures which is also influenced by the reader’s experience as a breast radiologist.”

8. Discussion page 19, line 10. 4.7% of women would be reallocated on cumulus. This is a very good agreement and better than many other studies using this technology. Any explanation for this?

**Answer:** Our correlation (ICC and Pearson) are comparable with others. The 5% of women who would have been allocated differently by the two readers are on a high/low-risk basis. In contrast to the BIRADS and Tabár classification the inter-observer discrepancy was random and the variance in PMD was not specifically pronounced in the proximity of the borderline (cut off) measures. This inclines toward a smaller high/low discrepancy compared to the other methods. One explanation why our high/low-risk agreement seems high could be contributed by the way we divide into quartiles. Instead of dividing into quartiles based on an equal number of controls in each group (which is often used when validating new methodologies with continuous data regarding the association with risk), we have instead divided into 4 groups with equal density.
range (more clinical thresholds corresponding to the “old” BIRADS density classification). This leads to a higher cut off value making the high/low-risk categorization less susceptible to the variance (which does not increase proportionally with increasing density). We have now also done the calculations based on the number of controls, and find that with these cut-offs 9.6% would have been allocated differently.

**Action:** We have given a more detailed definition of our categorization of the PMD measure.

**Methods p11 line 7-9:** “…(four groups with equal percentage density ranges within density range, corresponding to the BI-RADS classification).”

**Methods p12 line 12-14:** “PMD measured by the threshold technique was divided into four equal percentage ranges—quartiles within range of the PMD measures—corresponding to the BI-RADS categorization into density quartiles.”

**Reviewer 2: Jingmei Li**

**Reviewer's report:**

The authors compare inter-observer agreement and differences among three different methods of characterizing mammographic density - BIRADS, TABAR and area-based percent mammographic density. In addition to comparing risk by different categories, often by increasing density, the authors also combined categories into low/high risk, which decreased inter-observer variability. The conclusion is that the three methods generally predict breast cancer risk, but are associated with different effect sizes. The authors also highlighted that the Tabar patterns are associated with the highest discrimination of breast cancer risk in terms of odds ratios.
The question posed by the authors was well-defined, and the methods are extremely detailed and well-described. Although the parent screening study is huge (n=14736), the number of cases used for the breast cancer case-control study is relatively small (n=122).

**Discretionary Revisions**

1. Scatter plot and Bland Altman plot: the numbers on the axes have commas instead of "." to denote decimal points.

**Answer:** We thank the reviewer for pointing this out. We have used IBM SPSS Statistics 20, Copyright © IBM Corporation 1989-2011 for statistical analysis with the European default basic settings. This has now been corrected to US settings.

**Action:** Commas have been replaced by "." on the scatter and Bland Altman plot.
2. Do the results change if the DCIS cases were removed? Discuss the potential effect the DCIS cases would have on the effect sizes of breast cancer risk.

**Answer:** We appreciate the comment and agree with the reviewer that it would be interesting to include this discussion. However, the article is already very extensive with many areas being discussed. Moreover, our intention was to represent “real life” in the screening setting. With less than 10% of the cases being DCIS (representing “real life”) we find it beyond the scope of this article to include this discussion. We hope that the reviewers can agree.

3. **p14 lines 14 and 17.** The result is either significant or it isn’t. If the p-value is higher than 0.05, it should be called ‘non-significant’ and written up as such.

**Answer:** We are glad this wrong wording was pointed out and agree that it is not appropriate to write “nearly significant”.

**Action:** Page 14 lines 18 and 21. We have replaced “nearly significant” with “NS” when this appears throughout the manuscript. NS is specified under *List of abbreviations*

4. Apart from BMI, menopausal status is also strongly associated with breast density. The lack of information on menopausal status should be acknowledged too.

**Answer:** We agree with reviewer and kindly refer to our answer concerning the second comment by reviewer 3.

**Action:** We kindly refer to the changes made concerning the second comment by reviewer 3.
5. How does the risk prediction results look if you average the scores by both readers and use that in the model?

**Answer:** We appreciate this comment. We have done consensus scores on BIRADS and Tabar as well as averaged PMD scores. This shows the following ORs for BIRADS (D2:1.53 (NS), D3:2.35 and D4:4.16), PMD (Q2:1.84 (NS), Q3: 2.05, Q4:3.24) and Tabar (PI:1.82 (NS), PIII: 3.27, PIV: 4.45 and PV: 2.00 (NS)). These consensus scores are considered more accurate measures for the tree subjective methodologies, and will be used for comparison with two fully automated techniques (this is why we haven’t included them in this paper). We kindly refer to our answer to the first comment by reviewer 1.

6. **p20 line 17.** What is the overlap of women who are considered high risk among the three different methods? Are the high risk women always the same women, i.e. what is the proportion of women who are consistently judged to be in a high risk category?

**Answer:** We thank for this comment and agree with reviewer that this is an important and very interesting question. We are investigating this further in our ongoing study comparing five methods of which two are fully automated. In that study one of our main purposes is to compare density and texture scores to see if texture adds to density in estimating relative risk of breast cancer in terms of ORs. Even though the main purpose of the present study is to compare readers and not methods, we agree that it would be fine to elaborate briefly on the overlap of women considered high risk between methods.

**Action:** **Discussion: Page 21 line 1-4.** These sentences have been added: “In total only 23% of the women would consistently have been classified in the high-risk group by all three methods by R1.”
and 22% by R2. It is beyond the scope of this article to draw conclusions on, if this can partly be explained by the fact that the three methods may catch different risk parameters.”

Changes to the manuscript which are not on the request of reviewers:

1) The title has been changed from: “Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on risk prediction” to “Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer”.

2) After submission of the manuscript, images were rescanned on a new scanner to obtain better quality and resolution for the automated analyses for our next study. In this process, we recognised that one woman had entered the database with the wrong images. We have redone all analyses of data (including the corrected images). The recalculation...