Author’s response to reviews

Title: Texture analysis on MR images helps predicting non-response to NAC in breast cancer

Authors:

Nicolas Michoux (nicolas.michoux@uclouvain.be)
Stéphane van den Broeck (stephane.vandenbroeck@gmail.com)
Laure Lacoste (lacostelaure@hotmail.com)
Latifa Fellah (latifa.fellah@uclouvain.be)
Christine Galant (christine.galant@uclouvain.be)
Martine Berlière (martine.berliere@uclouvain.be)
Isabelle Leconte (isabelle.leconte@uclouvain.be)

Version: 9
Date: 18 June 2015

Author’s response to reviews: see over
Dear Dr Paolo Bruzzi,

It is my pleasure to hereby submit the revised version of our manuscript entitled "Texture analysis of MR images helps predicting non-response to NAC in breast cancer" for publication in BMC Cancer.

This letter is keen to provide a point-by-point response to all concerns made by reviewers 2 and 4, with a description of the changes made in the text. For all subsequent correspondence, please write me at the above address.

Yours Sincerely,

Nicolas Michoux, PhD.

Conflict of interest: none in respect to this work and its publication.
Financial interest: none in respect to this work and its publication.
All authors have agreed to the submission to the journal and that the manuscript is not currently under submission in any other journal.
**Detailed response**

______________________________ Reviewer 2

1. The authors have responded to most of my concerns. However, the measurement of residual tumor size in pathology in case of multiple foci should be further clarified.

As suggested, the description has been clarified and the text modified as follows: In case of a single mass lesion with a concentric response, the size of the residual tumor was measured. In case of a single mass lesion with a fragmented response, i) the overall dimension of the foci is given when foci are adjacent, ii) each foci is measured when foci are distant and the overall sum is given. In case of a non-mass lesion with fragmented response, the overall size is given.

______________________________ Reviewer 4

1. However, before they start with the analysis, they described the numerical variables with mean and standard deviation. The problem is, there was no justification for the use of these statisticians. It is not suitable to use them without being sure that the data was normally distributed, because the mean and the standard deviation are not an appropriate descriptor when data is not normally distributed. On the other hand, they mentioned that they couldn’t verify the normal distribution of the data and also used the D’Agostino-Pearson test to do so. There is no consistency in this matter.

As suggested, mean values and standard deviation have been replaced by median values and 95% confidence interval on the median (Table 4 and Text).

2. To compare the parameters of both tissues (healthy breast tissue vs tissue of malignant lesion) they decided to use the Wilcoxon rank sum test, because they could not verify the normal distribution of data. So why then, they described with mean and standard deviation? This is another inconsistency, because Wilcoxon compares medians not means.

See answer to comment 1.

3. No sample size calculation was made. So, what is the reliability of the differences found in these study, if the statistical tests used, were not based on the assumptions of the calculation of sample size? For example, in Fisher's test?

This study used retrospective materials. Due to the exploratory nature of the study (testing the theoretical hypothesis that a MRI model based on visual texture may help predict non response to NAC), no power analysis was performed (effect size being...
not known by the way). We agree that in the planning phase of a confirmatory study, sample size (based on power analysis) has to be set usefully to assess the previous exploratory hypothesis.

Let us note that results observed on biologic parameters agree with what is currently known (luminal A type tends to respond less to NAC, Ki67>14% and HR-/HER2+ are markers more often seen in responders).

4. Another point to consider is the difference between the sample size of the NR group (19) and the PR + CR group (50). To compare the groups, standardization was performed?

No weighting has been performed. Difference between sample sizes is expected to be removed within the confirmatory study.

5. When they compared with Fisher's exact test, biological parameters, BI-RADS, texture and kinetic, only one of these was significant luminal A. It would have been appropriate to report the performance of luminal A to predict non-response to NAC through the ROC curve and AUC.

Luminal A subgroup is small (13 patients, 8 NR, 0 CR, 5 PR) for a reliable assessment of the performance of the MRI model. Luminal A type is known to be less sensitive to NAC. Therefore, from a clinical point of view, it is more interesting to assess the performance in other types like triple-negative and luminal B [see refs 6-12].

6. They performed univariate predict models for each variable, where sensitivity, specificity, AUC and cut points were calculated. However, researchers did not explain what was the criterion used for the cut point. High sensitivity? Or high specificity? It is important to declare the selection criteria, because it defines the weight of each parameter to build the predictive model of non-response to NAC. Also defines the intentions of the researches, they want the model to be highly sensitive, or highly specific? This requirement also applies to the multivariate model.

The criterion used for cut-off values (ROC) and the identification of good predictive models, was the Youden Index, classically. As suggested, a comment on this has been added in the legend of Table 5.

7. The k-means clustering is not justified enough. The aim of this statistical method is to form grouping variables. It is an easy way to classify a given data set through a certain number of clusters. The authors used kmeans clustering as a classifier, to select the variables for the predicting model. However, this method does not classify the variables according to their association with the effect. In addition there is no control of the multicollinearity between the variables. The problem is that the interdependency may overestimate the model.
Coefficient of Determination, $R^2$ and report a major ability to predict the outcome of the study (non-response to NAC).

Indeed, in this context, $k$-means aims at identifying groups of patients similar to each other on certain characteristics (texture parameters). It does not classify the variables according to their association with the effect (NR or CR+PR), it performs a segmentation, looking for internal cohesion within groups and external isolation among groups; which is one way to address the problem of the evaluation of patient’s response to NAC according to MRI texture characteristics. The association with the effect is tested via the cross-validation + ROC analysis, performed for each clustering (as mentioned in page 11).

As rightly emphasized by reviewer 4, multicollinearity may be a problem in statistics; a problem that is periodically revisited since [Farrar DE, Multicollinearity in Regression Analysis: The Problem Revisited, The Review of Economics and Statistics 1967;49(1):92-107]. A cure to this problem is still under investigation [Dormann CF, Ecography 2013;36:027-046]. However, if interest is only in estimation and prediction as it is the case in the present study, multicollinearity may be ignored since it does not affect the fitted value or its standard error.

8. They also used the logistic regression as a classifier, nevertheless I think it could be a better way instead of doing a stepwise method. If the aim of the study was to identify the variables that have the greatest ability to predict the non-response to NAC, it could be better to use a method of selection variables that consider their relation with the effect and also control the multicollinearity between them.

There are different variables selection methods, the most common ones being the stepwise regression, information criterion, regularization and dimension reduction method. These methods should be used as a guide only, for identifying good predictors. They have limitations which have been summarized in [Harrell FE. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis, 2001, Springer-Verlag, New York] and in [Trevor Hastie RT. The elements of statistical learning, 2001, Springer-Verlag, New York].

A first selection of the most pertinent biologic, kinetic and texture parameters was performed according to experts (anapathologist, radiologist and physicist). As a result, only Ki67>14%, HR-/HER2+, the 3 most common kinetic parameters, only 9 Haralick’s texture parameters, and traditional run-length parameters were considered, with the aim to predict non-response to NAC (See answer to comment 7).

9. To fulfill these requirements, I suggest using the Principal Component Analysis (PCA). This method allows to identify and select the variables that give the greatest ability to predict the outcome (non-response to NAC). The virtues of this analysis are:
· Doesn’t require the variables to be normally distributed. The first step to classify the variables is the use of a correlation matrix, where the variables that have the best correlation with the outcome are selected.
· Considers the multicollinearity between variables in the model identifies them and choose the independent variables.
· The PCA hierarchically ordered the variables that give the higher variability of the model; this is given by the eigen value and the eigen vector (orthogonal) so it gives the assurance of independence and therefore, not collinearity. The PCA can identify the variables with greatest weight and effect on the outcome, in this case the non-response to NAC. Then, when the best variables are selected, you can perform a logistic regression that reports a determination coefficient (R2 adj), then a ROC curve with the AUC and calculate the sensitivity and specificity of the model.

We agree that many machine learning algorithms can be envisaged to assess the usability of texture parameters in breast MRI, with an influence on the diagnostic performance and generalization properties of the model. PCA, as suggested, is indeed an elegant approach to the problem of the evaluation of response to NAC, but also ICA or PLS-DA which have been used in other studies (cf. literature on MRI texture analysis for classification purpose), constitute potential candidates.

Given the task’s magnitude, only 2 classifiers, each belonging to one of the two classes of algorithms in machine learning (unsupervised and supervised), were tested. As a result, we mention in the discussion that further tests, in particular based on other learning models (but also using other set of 2D or 3D texture parameters) and relying on a new patient dataset, must be considered subsequently to this pilot study.