Turning subjective into objective: The BCCT.core software for evaluation of cosmetic results in breast cancer conservative treatment

Maria João Cardoso a,*, Jaime Cardoso b, Natália Amaral c, Isabel Azevedo d, Lise Barreau e, Mario Bernardo f, David Christie g, Susy Costa a, Florian Fitzal h, José L Fougo d, Jørgen Johansen i, Douglas Macmillan j, Maria Piera Mano k, Lea Regolo l, José Rosa l, Luís Teixeira b, Conny Vrieling m

a Department of Surgery, Faculdade de Medicina do Porto, Hospital S. João, Alameda do Prof. Hernâni Monteiro 4200-319, Porto, Portugal
b Unidade de Telecomunicações e Multimédia, INESC Porto, Porto, Portugal
c Hospital da Universidade de Coimbra, Coimbra, Portugal
d Instituto Português de Oncologia, Porto, Portugal
e Institut Gustave Roussy, Villejuif, France
f Instituto Português de Oncologia, Lisbon, Portugal
g East Coast Cancer Center, Tugun, Australia
h Medical University Vienna Waehringer, Austria
i Odense University Hospital, Odense, Denmark
j Nottingham Breast Institute, Nottingham City Hospital, Nottingham, UK
k University of Turin, CPO Piemonte, Italy
l Fondazione Salvatore Maugeri, Surgical Department, Pavia, Italy
m Netherlands Cancer Institute, Amsterdam, The Netherlands

Received 28 January 2007; received in revised form 22 March 2007; accepted 3 May 2007

KEYWORDS
Breast cancer; Conservative treatment; Aesthetic result; Expert observers; Nominal consensus;

Summary Twelve expert observers from nine different countries convened in a workshop to evaluate the validity of the Breast Cancer Conservative Treatment. Cosmetic results (BCCT.core) software, an objective method for the aesthetic evaluation of breast cancer conservative treatment.

Experts were initially asked to subjectively classify the aesthetic results of 30 photographed cases submitted to breast cancer conservative treatment according to the four-point Harris scale. It was pre-established that if at least two-thirds
[Cardoso MJ, Cardoso J, Santos AC, Barros H, Oliveira MC. Interobserver agreement and consensus over the esthetic evaluation of conservative treatment for breast cancer. Breast 2005] of participants provided the same classification this would be considered a consensual evaluation for that case. For cases where such agreement was not reached, consensus was obtained using a nominal group technique. Experts then individually performed objective evaluation of the same set of photographs using the BCCT.core software. This provides an automatic rating of aesthetic results, once scale and reference points in the photograph have been chosen. Agreement between observers, between each observer and the consensus, for computer evaluation obtained by the different participants and between software and consensus was calculated using multiple kappa ($k$) and weighted kappa ($wk$) statistics.

In the subjective assessment, first-round consensus was achieved in 17 (57%) cases. Overall interobserver agreement was fair to moderate ($k = 0.40$, $wk = 0.57$). In the objective assessment there was a higher level of concordance between participants ($k = 0.86$, $wk = 0.90$). Agreement between software and consensus classification was fair ($k = 0.34$, $wk = 0.53$), but was higher in the 17 cases that reached first-round consensus ($k = 0.60$, $wk = 0.73$). Merging the two middle classes of the Harris scale, to form a three-point scale, led to an improvement of all non-weighted measures of agreement.

These results show that the BCCT.core software provides consistent evaluation of cosmesis. It has the potential to become a gold standard method for assessment of breast cosmesis in clinical trials, as it can be used simultaneously by a panel of observers from different parts of the world to provide more reliable assessments than has been possible previously.
© 2007 Elsevier Ltd. All rights reserved.

**Introduction**

To obtain an acceptable cosmetic result is one of the major aims of breast cancer conservative treatment.$^{1,2}$ The absence of international standards of many treatment parameters, for example, the quantity of tissue to be excised around the tumor, limits the applicability of any comparative analysis of cosmetic outcome.$^3$ Although analysis of results for surgical procedures in terms of disease free survival and overall survival has become common practice,$^{1,2}$ assessment of cosmesis remains without a standard.$^{4,5}$ Methods for evaluating breast cancer conservative treatment are traditionally considered as subjective or objective.$^6$ Subjective methods usually evaluate a patient’s appearance on a photograph by observers and have frequently used personnel involved in the treatment process for this purpose. Personal experience in breast cancer conservative treatment seems to favor agreement over aesthetic evaluation.$^{4,5,7}$ However, results of subjective evaluation show only a modest interobserver agreement, even when performed by expert observers.$^8$ This methodology is also time consuming, and all these factors have probably contributed to the limited evaluation of aesthetic results in studies analyzing the outcome of breast cancer conservative treatment.$^{1,2}$ Objective methods use measurements taken from the patient or from photographs, and are based essentially on asymmetries between treated and non-treated breast.$^{9-11}$ These methods increase the reproducibility of assessment but it has been argued that they do not take into account the global appearance of aesthetic results, failing to include other aspects such as scar appearance and differences in color between breasts.$^{12}$

The Breast Cancer Conservative Treatment. cosmetic results (BCCT.core) software was developed to provide an evaluation of aesthetical results, not only from given measurements, but also from other parameters extracted from patients photographs.$^{13}$ The algorithms were incorporated into a software capable of automatically attributing an overall classification of aesthetical results, once scale and reference points have been chosen by the user. The aim was to develop a reproducible and widely available methodology for evaluation of aesthetic results in breast cancer conservative treatment, enabling effective comparison of outcome between centers.$^{14}$

The purpose of this workshop, was the international evaluation of the validity of the BCCT.core software by an invited panel of experts, acting as a comparison group for software classification.
Material and methods

Invitations to participate in the workshop were sent by email to 13 healthcare professionals directly involved in breast cancer conservative treatment, all of whom had participated in a previous consensus panel on classification of aesthetic results, and in that study obtained coincident answers with the final consensus classification in at least two-thirds of cases. Nine of these professionals accepted the invitation. Four were unable to attend and three of them suggested the participation of another health professional from the same institution, equally experienced in breast cancer treatment. A final total of 12 clinicians from nine different countries participated in the workshop. These were asked to individually evaluate a series of digital photographs taken from 30 women submitted to conservative breast cancer treatment (surgery and radiotherapy) at two different institutions. None of the workshop participants had previous contact with the cases. Treatment interventions had ended at least 1 year before photographs were taken. All patients signed an informed consent to participate in the study.

A digital camera with a resolution of at least 4 megapixels was used to take photographs in four positions: face arms down; face arms up; left side arms up; right side arms up. Images were copied to twelve individual computers and each of the observers was asked to classify the aesthetic results of all cases in one of four categories: excellent—treated breast nearly identical to untreated; good—treated breast slightly different from untreated; fair—treated breast clearly different from untreated but not seriously distorted; poor—treated breast seriously distorted.

The evaluation of each case was a priori considered consensual when at least two-thirds (≥8) of observers provided the same classification of aesthetic result. For the remaining cases, group appreciation and discussion was undertaken with statistic evaluation of group response and establishment of a consensus when at least two-thirds of participants agreed on the same classification.

Evaluation of the same set of photographs using the BCCT.core software was then carried out individually by each of the observers, to test the consistency of obtained results. Large differences in this parameter were not expected, as the software performs automatic assignment to a given class, once scale and reference points in the photograph have been selected by the user (Fig. 1). Observers that did not conclude software evaluation of the assigned cases were excluded from this part of the study.

Agreement between observers, between each observer and the consensus, between each expert’s evaluation by using the software, and between software and consensus was evaluated by the multiple kappa (κ) and weighted kappa (wk) statistics, the latter allowing some deviation from perfect agreement. A kappa score equal to 0 was considered to indicate poor agreement; 0.01–0.20 slight agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 substantial agreement; 0.81–0.99 almost perfect; and 1.00 perfect agreement.
Results

Out of the 30 cases, 17 (57%) were attributed the same Harris class by at least two-thirds of observers on individual evaluation, and were therefore immediately considered consensual (first-round consensus). The remaining 13 cases were attributed the same classification by at least two-thirds of the panel after open discussion.

Main results of the agreement analysis are considered in Table 1. Overall interobserver agreement was fair to moderate and slightly higher for the 17 first-round consensus cases. Median agreement between experts and consensus was similar to interobserver agreement. Computer evaluation was very consistent when performed by the workshop participants. However, three clinicians failed to complete the analysis of the 30 study cases (did not save all the software answers).

Agreement between computer and consensus was moderate to substantial in the first 17 first-round cases, but was only fair overall.

Because a large number of disagreements were observed in the two middle classes of the Harris scale, results were recalculated after merging these two classes to form a modified three-point Harris scale (Table 2). This led to an improvement in all non-weighted measures of agreement, and substantial agreement was obtained between the median expert and consensus. Agreement between software and consensus reached a similar value.

Table 1 Evaluation in four classes: agreement between subjective (expert panel) results and objective (software) results.

<table>
<thead>
<tr>
<th>Subjective results</th>
<th>Objective results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First-round</td>
</tr>
<tr>
<td>Number of patients</td>
<td>17</td>
</tr>
<tr>
<td>Number of experts</td>
<td>12</td>
</tr>
<tr>
<td>Interobserver agreement (k; wk)</td>
<td>0.58; 0.73</td>
</tr>
<tr>
<td>Expert with highest agreement with consensus (k; wk)</td>
<td>0.84; 0.91 (2 differences)</td>
</tr>
<tr>
<td>Expert with lowest agreement with consensus (k; wk)</td>
<td>0.52; 0.67 (7 differences)</td>
</tr>
<tr>
<td>Expert with median agreement with consensus (k; wk)</td>
<td>0.72; 0.83 (4 differences)</td>
</tr>
<tr>
<td>Agreement between software and consensus (k; wk)</td>
<td>0.60; 0.73 (5 differences)</td>
</tr>
</tbody>
</table>

Discussion

The limited reproducibility of subjective aesthetic evaluation in breast cancer conservative treatment has been well documented and is one of the main reasons behind the development of objective methods for this purpose. The latter usually produces extremely reproducible results, but their application raises the problem of what is the standard to be used for comparison, given the limitations of subjective analysis. Comparison with panel evaluation seems the most widely used method but this study suggests that consensus may be easier in some cases (first-round consensus) than in others. Agreement is also very different in these two groups of cases. So, agreement between computer and consensus in the group of cases where experts reach an easier agreement is probably more relevant to the validity of the method than in those cases where consensus was difficult to attain.

Another difficult question is when is agreement between computer and consensus considered good enough for the first to substitute the subjective evaluation? Is almost perfect agreement with the consensus needed? The study of agreement between each of the observers and consensus helps us to put some perspective into this issue. Even the expert having the highest agreement with the consensus, only reached substantial agreement in the four-point scale (k = 0.73) although almost perfect agreement in the modified three-point scale (k = 0.87). It should be noticed that nine out of the 12 experts were selected from a very
experienced group of clinicians, having obtained the best agreement with consensus in a previous evaluation panel.\textsuperscript{8} We suggest that these are the agreement results to aim for in the development of objective computer analysis.

It is well known that the number of classes used for classification reflects strongly on agreement results. In aesthetic evaluation of breast cancer conservative treatment the four-point Harris scale has been traditionally used, with relatively poor agreement results.\textsuperscript{18} Similar findings were obtained in this study. But are the four-classes in the Harris scale necessary? Pezner et al.\textsuperscript{18} showed that changing from a four-point scale to a two-point scale in a study of 44 observers evaluating 14 photographs doubled the value of consensual answers ($k = 0.4 - 0.8$). In our study, reducing from a four-point to a three-point scale increased overall $k$ from 0.40 to 0.51.

The adoption of a lower number of classes for evaluation of aesthetic results of breast cancer conservative treatment seems a logical necessity, given the limited reproducibility of the four-point Harris scale. Much of the discussion arising in the workshop derived from the difficulty in agreement over classification in the two middle classes, suggesting that subjective discrimination between these categories is poor. Similar difficulties were encountered in the development of the computer classification algorithms.

Other authors have compared subjective with objective evaluation of breast cancer conservative treatment. Pezner et al.\textsuperscript{9} and Van Limbergen et al.\textsuperscript{11} introduced asymmetry measurements as the first form of objective evaluation aiming at more consistent results. They both concluded that quantitative assessments correlate well with subjective scoring, making this method relevant for clinical application. Since the publication of asymmetry measurements for evaluating aesthetic results it became a generalized thought that this simple form of objective assessment was very reproducible although not yet capable of traducing all aspects of cosmesis.

The use of BCCT.core software by different clinicians in the workshop showed that evaluation can be obtained consistently by users with no previous experience with the program, suggesting that it can be applied on a large scale.

Agreement between the software and consensus reached promising results, namely regarding comparison with first-round consensus in the three-point scale ($k = 0.72$, substantial agreement, three differences). Overall agreement with consensus was similar to that of the median expert.

The obtained analysis will be used to optimize the software, in view of the envisioned goal, obtaining an evaluation which is as close to the consensus as the best of the experts.

### Conflicts of interest

The first two authors are the main researchers involved in the development of the BCCT.core software. No commercial conflicts of interest exist.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Evaluation in three classes (good and fair merged): agreement between subjective (expert panel) results and objective (software) results.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subjective results</strong></td>
<td><strong>Objective results</strong></td>
</tr>
<tr>
<td>First-round consensus</td>
<td>Overall</td>
</tr>
<tr>
<td>Number of patients</td>
<td>24</td>
</tr>
<tr>
<td>Number of experts</td>
<td>12</td>
</tr>
<tr>
<td>Interobserver agreement ($k; wk$)</td>
<td>0.61; 0.66</td>
</tr>
<tr>
<td>Expert with highest agreement with consensus ($k; wk$)</td>
<td>1.00; 1.00 (0 differences)</td>
</tr>
<tr>
<td>Expert with lowest agreement with consensus ($k; wk$)</td>
<td>0.49; 0.56 (7 differences)</td>
</tr>
<tr>
<td>Expert with median agreement with consensus ($k; wk$)</td>
<td>0.77; 0.79 (3 differences)</td>
</tr>
<tr>
<td>Agreement between software and consensus ($k; wk$)</td>
<td>0.72; 0.79 (3 differences)</td>
</tr>
</tbody>
</table>
Acknowledgments

The authors would like to thank University of Porto Innovation Office for the organization of the meeting. The authors would also like to thank AstraZeneca, Calouste Gulbenkian Foundation, Champalimaud Foundation, Novartis Oncology, Roche, Sanofi Aventis and Science and Technology Foundation for funding expenses related to the BCCT.core workshop.

The authors also wish to thank Prof. Ayres-de-Campos for his help in preparation of the manuscript.

References