The Breast (2007) **16**, 456–461

ELSEVIER

THE BREAST

www.elsevier.com/locate/breast

ORIGINAL ARTICLE

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[f], 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

[*] Corresponding author. Tel.: +351 966484826; fax: +351 22 5505589.
E-mail address: mjcard@mail.med.up.pt (M.J. Cardoso).

doi:10.1016/j.breast.2007.05.002

Objective evaluation;
Software

[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 achived 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.

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.

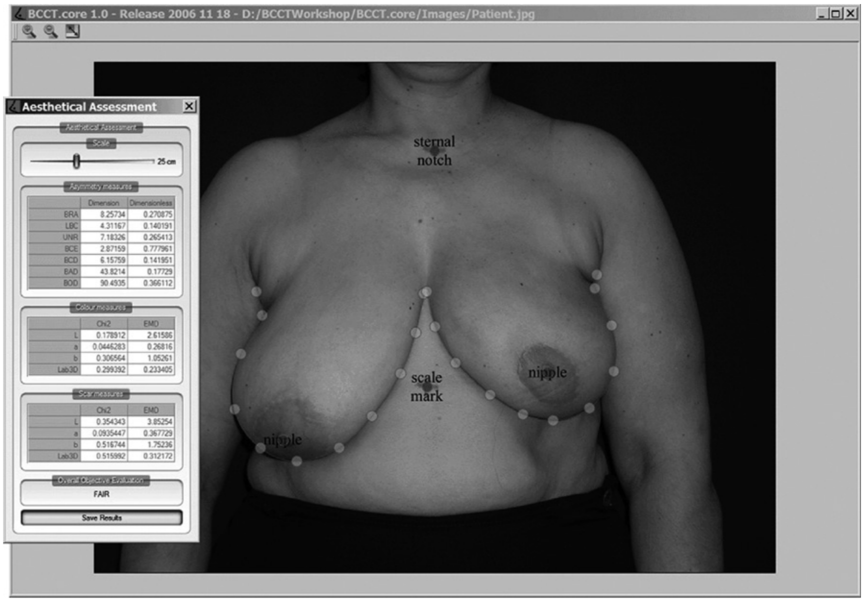
Material and methods

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.[8] 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[15]: excellent—treated breast nearly identical to untreated breast; 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 ($\geqslant 8$) 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[16] 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 (*k*) 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; 0.610.80 substantial agreement; 0.810.99 almost perfect; and 1.00 perfect agreement.[17]



BCCT.core 1.0 - Release 2006 11 18 - D:/BCCTWorkshop/BCCT.core/Images/Patient.jpg

Aesthetical Assessment

Aesthetical Assessment

Scale: 25 cm

Asymmetry measures

	Dimension	Dimensionless
BRA	8.25734	0.270875
LBC	4.31167	0.140191
UNR	7.18326	0.265413
BCE	2.87159	0.777961
BCD	6.15759	0.141951
BAD	43.8214	0.17729
BOD	90.4935	0.366112

Colour measures

	Chi2	EMD
L	0.178912	2.61586
a	0.0446283	0.26816
b	0.306564	1.05261
Lab3D	0.299392	0.233405

Scar measures

	Chi2	EMD
L	0.354343	3.85254
a	0.0935447	0.367729
b	0.516744	1.75236
Lab3D	0.515992	0.312172

Overall Objective Evaluation: FAIR

Figure 1 Scale and reference points in the photograph.

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.

Discussion

The limited reproducibility of subjective aesthetic evaluation in breast cancer conservative treatment has been well documented[8,18] and is one of the main reasons behind the development of objective methods for this purpose. The latter usually produces extremely reproducible results,[9–11] 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[4–6,19] 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

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

	Subjective results		Objective results	
	First-round consensus	Overall	First-round consensus	Overall
Number of patients	17	30	17	30
Number of experts	12	12	10	9
Interobserver agreement (*k*; *wk*)	0.58; 0.73	0.40; 0.57	0.75; 0.83	0.86; 0.90
Expert with highest agreement with consensus (*k*; *wk*)	0.84; 0.91 (2 differences)	0.73; 0.82 (6 differences)		
Expert with lowest agreement with consensus (*k*; *wk*)	0.52; 0.67 (7 differences)	0.37; 0.58 (14 differences)		
Expert with median agreement with consensus (*k*; *wk*)	0.72; 0.83 (4 differences)	0.57; 0.70 (10 differences)		
Agreement between software and consensus (*k*; *wk*)			0.60; 0.73 (5 differences)	0.34; 0.53 (14 differences)

Table 2 Evaluation in three classes (good and fair merged): agreement between subjective (expert panel) results and objective (software) results.

	Subjective results		Objective results	
	First-round consensus	Overall	First-round consensus	Overall
Number of patients	24	30	24	30
Number of experts	12	12	9	9
Interobserver agreement (*k*; *wk*)	0.61; 0.66	0.51; 0.57	0.82; 0.83	0.87; 0.88
Expert with highest agreement with consensus (*k*; *wk*)	1.00; 1.00 (0 differences)	0.87; 0.88 (2 differences)		
Expert with lowest agreement with consensus (*k*; *wk*)	0.49; 0.56 (7 differences)	0.40; 0.47 (11 differences)		
Expert with median agreement with consensus (*k*; *wk*)	0.77; 0.79 (3 differences)	0.62; 0.66 (6 differences)		
Agreement between software and consensus (*k*; *wk*)			0.72; 0.79 (3 differences)	0.57; 0.61 (6 differences)

experienced group of clinicians, having obtained the best agreement with consensus in a previous evaluation panel.[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.[18] Similar findings were obtained in this study. But are the four-classes in the Harris scale necessary? Pezner et al.[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.[9] and Van Limbergen et al.[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.

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

1. Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *N Engl J Med* 2002;**347**(16):1233–41.
2. Veronesi U, Cascinelli N, Mariani L, Greco M, Saccozzi R, Luini A, et al. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *N Engl J Med* 2002;**347**(16):1227–32.
3. Christiaens MR, van der Schueren E, Vantongelen K. More detailed documentation of operative procedures in breast conserving treatment: what good will it do us? *Eur J Surg Oncol* 1996;**22**(4):326–30.
4. Christie D, O'Brien M, Christie J, Kron T, Ferguson S, Hamilton C, et al. A comparison of methods of cosmetic assessment in breast conservation treatment. *The Breast* 1996;**5**:358–67.
5. Vrieling C, Collette L, Bartelink E, Borger JH, Brenninkmeyer SJ, Horiot JC, et al. Validation of the methods of cosmetic assessment after breast-conserving therapy in the EORTC "boost versus no boost" trial. EORTC radiotherapy and breast cancer cooperative groups. European organization for research and treatment of cancer. *Int J Radiat Oncol Biol Phys* 1999;**45**(3):667–76.
6. Al-Ghazal SK, Blamey RW. Cosmetic assessment of breast-conserving surgery for primary breast cancer. *Breast* 1999;**8**(4):162–8.
7. Cardoso MJ, Santos AC, Cardoso J, Barros H, Oliveira MC. Choosing observers for the evaluation of aesthetic results in breast cancer conservative treatment. *Int J Radiat Oncol Biol Phys* 2005;**61**(3):879–81.
8. 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;**1**:12.
9. Pezner RD, Patterson MP, Hill LR, Vora N, Desai KR, Archambeau JO, et al. Breast retraction assessment: an objective evaluation of cosmetic results of patients treated conservatively for breast cancer. *Int J Radiat Oncol Biol Phys* 1985;**11**(3):575–8.
10. Tsouskas LI, Fentiman IS. Breast compliance: a new method for evaluation of cosmetic outcome after conservative treatment of early breast cancer. *Breast Cancer Res Treat* 1990;**15**(3):185–90.
11. Van Limbergen E, van der Schueren E, Van Tongelen K. Cosmetic evaluation of breast conserving treatment for mammary cancer. 1. Proposal of a quantitative scoring system. *Radiother Oncol* 1989;**16**(3):159–67.
12. Triedman SA, Osteen R, Harris JR. Factors influencing cosmetic outcome of conservative surgery and radiotherapy for breast cancer. *Surg Clin North Am* 1990; **70**(4):901–16.
13. Cardoso JS, Pinto da Costa JF, Cardoso MJ. Modelling ordinal relations with SVMs: An application to objective aesthetic evaluation of breast cancer conservative treatment. *Neural Networks* 2005;**18**(5–6):808–17.
14. Cardoso J, Cardoso M. Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. *Artificial Intelligence in Medicine* 2007;**40**(2): 115–26.
15. Harris JR, Levene MB, Svensson G, Hellman S. Analysis of cosmetic results following primary radiation therapy for stages I and II carcinoma of the breast. *Int J Radiat Oncol Biol Phys* 1979;**5**(2):257–61.
16. Jones J, Hunter D. Consensus methods for medical and health services research. *Brit Med J* 1995;**311**(7001): 376–80.
17. Seigel DG, Podgor MJ, Remaley NA. Acceptable values of kappa for comparison of two groups. *Am J Epidemiol* 1992;**135**(5):571–8.
18. Pezner RD, Lipsett JA, Vora NL, Desai KR. Limited usefulness of observer-based cosmesis scales employed to evaluate patients treated conservatively for breast cancer. *Int J Radiat Oncol Biol Phys* 1985;**11**(6):1117–9.
19. Sacchini V, Luini A, Tana S, Lozza L, Galimberti V, Merson M, et al. Quantitative and qualitative cosmetic evaluation after conservative treatment for breast cancer. *Eur J Cancer* 1991;**27**(11):1395–400.