The Breast (2006) **15**, 52–57

ELSEVIER

THE BREAST

www.elsevier.com/locate/breast

ORIGINAL ARTICLE

Interobserver agreement and consensus over the esthetic evaluation of conservative treatment for breast cancer

Maria João Cardoso[a,*], Jaime Cardoso[b], Ana Cristina Santos[c], Henrique Barros[c], Manuel Cardoso de Oliveira[a]

[a]*Serviço de Cirurgia B, H.S. João, Faculdade de Medicina do Porto, Alameda do Prof. Hernâni Monteiro, 4200-319 Porto, Portugal*
[b]*Unidade de Telecomunicações e Multimédia, INESC, Porto, Portugal*
[c]*Serviço de Higiene e Epidemiologia, Faculdade de Medicina do Porto, Porto, Portugal*

Received 2 December 2004; received in revised form 14 March 2005; accepted 8 April 2005

KEYWORDS
Breast cancer;
Conservative treatment;
Esthetic result;
Expert observers;
Interobserver agreement;
Consensus;
Delphi approach

Summary Twenty-four experts from 13 different countries were asked to evaluate photographs taken of 60 women following conservative breast cancer treatment. The esthetic result of each case was classified as poor, fair, good or excellent. Agreement was evaluated using the kappa (k) and weighted kappa (wk) statistics, for all observers, male and female participants, those younger and older than 50 years, those seeing more than 250 cases a year, and those with previous publications in this area. Consensus was obtained by way of a modified Delphi approach, when more than 50% of participants provided the same classification. In a second round, consensual cases were disclosed and a revised opinion was asked in non-consensual ones. Agreement between all participants was fair ($k = 0.24$, wk $= 0.37$) and remained within the same range ($k = 0.20-0.31$, wk $= 0.31-0.45$) in the subgroups analyzed. First round consensus was obtained in 46 out of 60 cases (77%) and in the second round in 59 out of 60 cases (98%). Evaluation of the esthetic results of conservative treatment for breast cancer is only fairly reproducible when performed by experts working in different geographical areas. Consensus is obtainable if a relatively low threshold of agreement is considered acceptable.

Introduction

A good cosmetic result is an important endpoint for the conservative treatment of breast cancer, but the verification of this outcome remains without a

*Corresponding author. Tel./fax: +351 225505589.
E-mail address: mjcard@mail.med.up.pt (M.J. Cardoso).

0960-9776/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.breast.2005.04.013

standard.[1,2] Methods of evaluating the results of conservative treatment for breast cancer are traditionally considered to be of an objective[3–5] or a subjective nature.[6–9] Objective methods use measurements taken from the patient or from photographs and are based on asymmetries presented by the treated breast compared with those of the non treated one.[3,4,10] They seem to be highly reproducible, but it could be argued that they do not take into account the overall appearance of the individual.[11] Subjective methods imply observer evaluation of the patient's appearance after treatment, and have usually used personnel involved in the treatment process[8,9] or the patients themselves.[12] It is common for observers to be recruited from the local medical or non-medical staff, for practical reasons.[7,13] Personal experience of the conservative treatment of breast cancer seems to favor agreement,[2,13,14] perhaps because the general esthetic features valued by society are not confused with aspects related to the surgical procedure itself.

The aim of this study was to evaluate interobserver agreement over the esthetical results of conservative treatment for breast cancer by a large panel of experts in this field working in different geographical areas. It was hoped that this methodology would also allow a widely achieved consensus on the evaluation of these cases, thereby providing a ''standard'' to be used for the optimization of more objective and quantifiable methods.

Materials and methods

Invitations to participate in the study as observers were sent by email to 40 professionals with previous experience of breast cancer diagnosis and treatment, working in 20 different countries. International recognition, number of cases seen or treated per year, and previous work published on the esthetic results of conservative treatment for breast cancer were the criteria used for invitation. Twenty-four experts from 13 different countries agreed to participate (Table 1). They were asked to individually evaluate a series of digital photographs taken from 60 women who had undergone conservative breast cancer treatment (surgery and radiotherapy). Treatment interventions had ended at least one year before the patients were photographed. All patients signed an informed consent to participate in the study. A digital camera with a resolution of four mega pixels was used, with a blue panel as a background. Photographs were taken in four positions, with the patient standing on floor marks: facing with arms down; facing with arms up; from the left side with arms up; and from the right side with arms up (Fig. 1). Images were recorded on a CD and posted to all observers, who were anonymous to each other, with detailed instructions on how to proceed with analysis and a request to return the answers, by email, as quickly as possible. Participants were asked to evaluate individually the esthetic result in each case, classifying it into one of four categories:[6] 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.

Agreement between observers 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; 0.61–0.80 substantial agreement; 0.81–0.99 almost perfect; and 1.00 perfect agreement.[15]

Agreement was evaluated for all observers and also for the following subgroups: male and female participants, observers aged more or less than 50 years, those who had seen more or fewer than 250 cases of conservative treatment for breast cancer per year, and authors of papers on the esthetic results of conservative treatment for breast cancer.

In order to obtain a consensus, the Delphi process, with the conducting of several rounds of agreement, was used.[16,17] The evaluation of each case was considered consensual when more than 50% of observers provided the same classification of esthetic result. For subsequent rounds, feedback sheets of previous results were sent by email, disclosing consensual cases and asking for a revised opinion on non-consensual ones. Interobserver agreement was analyzed for each round, according to the previously described methodology. Again, the evaluation of each case was considered consensual when more than 50% of observers provided the same evaluation.

Results

Agreement among all experts and in the different subgroups is displayed in Table 2. Results obtained with the k statistic indicate a fair agreement in all subgroups of analysis. The wk statistics provided slightly higher values, but failed to exceed 0.50 in any group of observers. Consensus in the first round occurred in 46 out of 60 cases (77%). In the second

Table 1 Main characteristics of expert observers.

Gender	Age	Working area	Number of patients/year	Publication in this field	Country
Male	⩾50	Breast surgeon/clinical oncology	>500	No	Italy
Female	<50	Radiation oncology	<100	Yes	Netherlands
Male	<50	Radiation oncology	70	Yes	Australia
Male	<50	Breast surgeon	>500	Yes	Netherlands
Male	<50	Breast surgeon	>100<250	No	Portugal
Male	⩾50	Radiologist	>100<250	No	Portugal
Female	<50	Breast surgeon/ reconstructive surgery	>250<500	No	UK
Male	>50	Breast surgeon	150	Yes	Sweden
Female	<50	Radiation oncology	>250<500	Yes	Portugal
Female	>50	Radiation oncology	>100<250	No	Portugal
Male	<50	Surgeon, consultant	>500	No	Slovenia
Male	<50	Radiation/medical oncology	<100	Yes	Denmark
Male	⩾50	Reconstructive surgery	<100	No	Portugal
Female	⩾50	Radiation oncology	<100	Yes	USA
Female	⩾50	Reconstructive breast surgeon	>250<500	No	France
Female	⩾50	Breast surgeon/ gynecology	>250<500	Yes	Italy
Male	<50	Breast surgeon	>250<500	No	UK
Male	⩾50	Breast surgeon	>250<500	No	Spain
Male	⩾50	Breast surgeon/ gynecology	>500	No	Portugal
Male	⩾50	Breast surgeon	>500	No	Austria
Female	⩾50	Breast surgeon	>100<250	Yes	Finland
Male	<50	Radiologist	>250<500	No	UK
Female	⩾50	Medical oncology	>100<250	No	Portugal
Male	<50	Breast surgeon	>100<250	Yes	USA

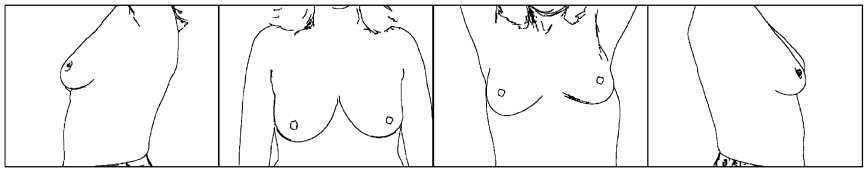

Figure 1 Positions used for the photographs.

round, answers were obtained from 22 of the 24 participants and consensus was reached in 59 out of 60 cases (98%). Given this result, it was considered unnecessary to proceed with further rounds of consensus.

Of the 22 participants evaluating the remaining 14 cases of the second round, three did not alter any of their evaluations. In 76 cases, participants changed their evaluation into a contiguous category. In 11 cases this change occurred between non-contiguous categories.

Discussion

Agreement on the subjective evaluation of the esthetic results of the conservative treatment of breast cancer has been previously reported by others. Vrieling et al.[2] asked five observers to evaluate the results of 731 patients according to the four-point classification system used in the present study, and reported a fair interobserver agreement ($k = 0.28$, wk = 0.42). Sneeuw et al.[18] assessed the evaluation of 76 patients by two

Table 2 Interobserver agreement in classification of esthetic results.

Groups	1st round			2nd round		
	n	*k*	wk	*n*	*k*	wk
All observers	24	0.24	0.37	22	0.30	0.43
Male	15	0.22	0.36	14	0.28	0.41
Female	9	0.29	0.41	8	0.33	0.46
$<$50 years	11	0.29	0.44	10	0.34	0.48
$\geqslant$50 years	13	0.21	0.31	12	0.26	0.39
$<$250 cases a year	12	0.25	0.38	12	0.28	0.42
$\geqslant$250 cases a year	12	0.23	0.36	10	0.30	0.45
Publications	10	0.31	0.45	10	0.35	0.49
No publications	14	0.20	0.31	12	0.24	0.37

n—number of observers; *k*—multiple kappa statistic; wk—weighted kappa statistic

experienced observers using the four-point system and found a high interobserver agreement ($k = 0.64$). Christie et al.[13] studied the assessment of 47 photographs by five observers (two trained, three untrained), again using the four-point classification system. They report an absolute agreement in 49% of cases assessed by trained personnel and 19% by untrained observers. We have previously reported on the evaluation of 55 cases by 13 observers (four experienced, four moderately experienced, and five inexperienced). Agreement was higher in the group of experienced observers ($k = 0.59$) than in the moderately experienced ($k = 0.35$) and inexperienced observers ($k = 0.33$).[14] The latter group of observers comprised non-clinical personnel from the community (mathematicians, nutritionists, etc.). Thus, while it seems intuitive that the opinion of non-specialized people is closer to the real evaluation of results by society, their assessment is much less reproducible than that of experts. It is possible that this is because the non-specialists have difficulties in separating general characteristics of the breasts from the symmetry parameters, color differences, and scar appearance related to the esthetic consequences of this form of treatment.

Until now, evaluation of agreement between experts has been restrained to observers working in the same institution or in rare cases the same or close countries. Experts working in different geographical areas have not been involved, probably because of practical and financial constraints. The current reality of widespread use of digital photographs, computer imaging programs, and email communication has made this evaluation possible at a relatively low cost. Our results show that evaluation of the esthetic results of conservative treatment for breast cancer, when performed by experienced individuals working in different geographical areas, has a limited reproducibility. Interobserver agreement remains fair when performed by male or female participants, by those younger or older than 50 years, by those seeing more or fewer than 250 cases a year, and by those with or without previous publications in this area. This suggests that there is considerable variation in evaluation of the esthetic results of conservative breast cancer treatment in different parts of the world.

The four-category scoring system used in this study, first described by Harris et al. in 1979,[6] is widely used in reproducibility studies,[2,13,14] but it is possible that if fewer categories are used, agreement will be higher. However, the values obtained with the wk statistic, where assignment to adjacent categories has a less negative impact on results, suggest that improvements would not be considerable.

The Delphi approach was used in order to establish a consensus, a methodology that has no rigid pre-established rules, and that needs to be adapted to the situation under evaluation.[16] A universally established percentage of agreement for obtaining a consensus does not exist for the Delphi approach.[17] The option taken in this study was for a relatively low value of agreement (50%) because of the limited reproducibility obtained in the first round, and the fear that an unacceptable number of rounds would be necessary with a consequently high dropout rate. In fact, the pre-established time limit to return the answers had to be postponed because of delays from several participants. The high workload involved for observers participating in this study (in the first round it implied the opening and evaluation of 240 image files on the computer) raised the concern that there could be many dropouts for subsequent rounds. Other authors have reported on the

difficulty of maintaining high levels of participation in similar studies.[19] As it was, these concerns were probably exaggerated, as only two of the 24 panelists dropped out in the second round and consensus in all but one case was reached at the end of this round.

Poor reproducibility is a problem that affects many aspects of medical care, and is the main motivation behind the development of objective and/or computerized methods of evaluation.[20] However, the assessment of these methods needs a "standard" by which to compare results, and for this there is usually no alternative to subjective evaluation by observers. Hence, the poor reproducibility documented but the need for a consensus over esthetic evaluation of the results of conservative treatment for breast cancer.

In conclusion, the esthetic evaluation of the results of conservative treatment for breast cancer by experts working in different geographical areas has a limited reproducibility. A consensus on this evaluation, however, can be achieved if a relatively low threshold of agreement is considered acceptable. At the close of this study, a consensual interpretation of the results of 59 cases of conservatively treated breast cancer is available as a "standard" by which to compare objective methodologies.

Acknowledgments

We wish to thank Prof. Ayres-de-Campos for his help in preparation of the manuscript, and to acknowledge the observers who kindly agreed to participate, and who dedicated their time to the evaluation of the cases, providing important feedback for this paper.

The following were participants in the study: Bruno Salvadori, (Policlinica S. Marco, Bergamo, Italy), Conny Vrieling (Netherlands Cancer Institute, Amsterdam, The Netherlands), David Christie (East Coast Cancer Center, Tugun, Australia), Emiel Rutgers (Netherlands Cancer Institute, Amsterdam, The Netherlands), Fernando Castro (Instituto Português de Oncologia, Porto, Portugal), Fernando Lage (Instituto Português de Oncologia, Lisbon, Portugal), Fiona MacNeill (Essex County Hospital, Colchester, UK), Goran Liljegren (University Hospital, Örebro, Sweden), Isabel Azevedo (Instituto Português de Oncologia, Porto, Portugal), Isabel Monteiro Grillo (Hospital Santa Maria, Lisbon, Portugal), Janez Zgajnar (Institute of Oncology, Ljubljana, Slovenia), Jørgen Johansen (University Hospital, Odense, Denmark), José Rosa (Instituto Português de Oncologia, Lisbon, Portugal), Leela Krishnan (University of Kansas Medical Center, Kansas, USA), Lise Barreau (Institut Gustave Roussy, Villejuif, France), Maria Piera Mano (CPO Piemonte, Torino, Italy), Michael Dixon (Edinburgh Breast Unit, Western General Hospital, Edinburgh, Scotland, UK), Miguel Prats Esteve (University Hospital, Barcelona, Spain), Natália Amaral (Hospital da Universidade, Coimbra, Portugal), Raimund Jakesz (AKH University of Vienna, Austria), Rauni Saaristo (University Hospital, Tampere, Finland), Robin Wilson (Nottingham Breast Institute, City Hospital, Nottingham, UK), Vera Tomé (Instituto Português de Oncologia, Lisbon, Portugal), Virgilio Sacchini (Memorial Sloan Ketterin Cancer Center, New York, USA).

References

1. Al-Ghazal SK, Blamey RW. Cosmetic assessment of breast-conserving surgery for primary breast cancer. *Breast* 1999;**8**(4):162–8.
2. Vrieling C, Collette L, Bartelink E, 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.
3. Pezner RD, Patterson MP, Hill LR, 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.
4. 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.
5. Krishnan L, Stanton AL, Collins CA, Liston VE, Jewell WR. Form or function? Part 2. Objective cosmetic and functional correlates of quality of life in women treated with breast-conserving surgical procedures and radiotherapy. *Cancer* 2001;**91**(12):2282–7.
6. 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.
7. Beadle GF, Silver B, Botnick L, Hellman S, Harris JR. Cosmetic results following primary radiation therapy for early breast cancer. *Cancer* 1984;**54**(12):2911–8.
8. Liljegren G, Holmberg L, Westman G. The cosmetic outcome in early breast cancer treated with sector resection with or without radiotherapy. Uppsala-Orebro Breast Cancer Study Group. *Eur J Cancer* 1993;**29A**(15):2083–9.
9. Cetintas S. Factors influencing cosmetic results after breast conserving management (Turkish experience). *Breast* 2002;**11**(1):72–80.
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. Triedman SA, Osteen R, Harris JR. Factors influencing cosmetic outcome of conservative surgery and radiotherapy for breast cancer. *Surg Clin N Am* 1990;**70**(4):901–16.

12. Al-Ghazal SK, Fallowfield L, Blamey RW. Patient evaluation of cosmetic outcome after conserving surgery for treatment of primary breast cancer. *Eur J Surg Oncol* 1999;**25**(4): 344–6.
13. Christie DR, O'Brien MY, Christie TK, et al. A comparison of methods of cosmetic assessment in breast conservation treatment. *Breast* 1996;**5**:358–67.
14. 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.
15. Seigel DG, Podgor MJ, Remaley NA. Acceptable values of kappa for comparison of two groups. *Am J Epidemiol* 1992;**135**(5):571–8.
16. Jones J, Hunter D. Consensus methods for medical and health services research. *Br Med J* 1995;**311**(7001):376–80.
17. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs* 2000;**32**(4):1008–15.
18. Sneeuw KC, Aaronson NK, Yarnold JR, et al. Cosmetic and functional outcomes of breast conserving treatment for early stage breast cancer. 1. Comparison of patients' ratings, observers' ratings, and objective assessments. *Radiother Oncol* 1992;**25**(3):153–9.
19. Powell C. The Delphi technique: myths and realities. *J Adv Nurs* 2003;**41**(4):376–82.
20. Ayres-de-Campos D, Bernardes J, Costa-Pereira A, Pereira-Leite L. Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision. *Br J Obstet Gynaecol* 1999;**106**(12):1307–10.