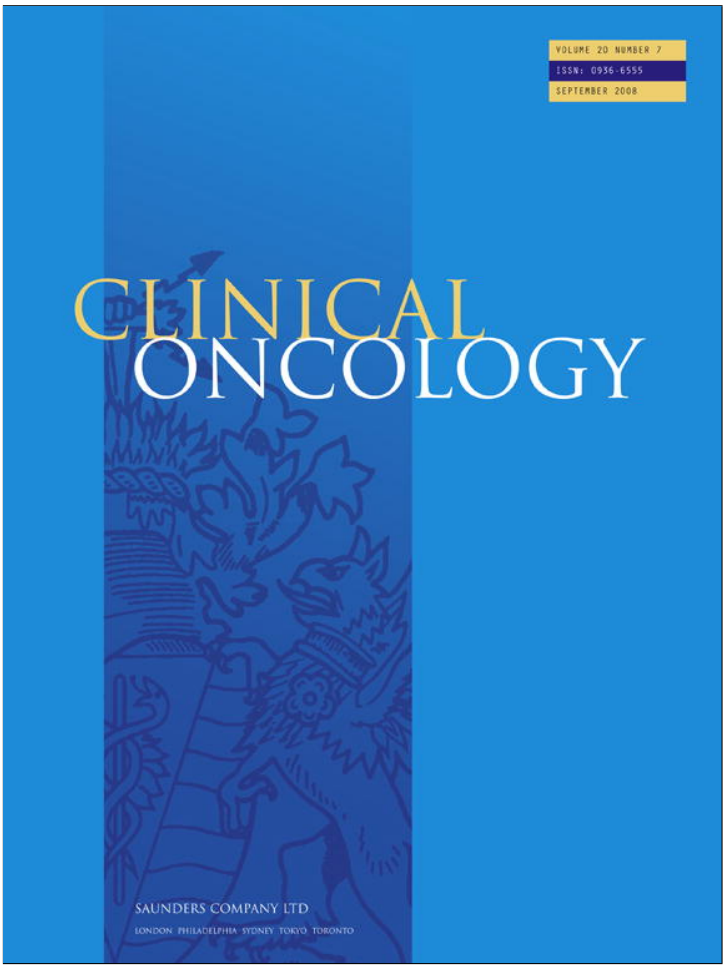

Clinical Oncology (2008) **20:** 497–501
doi:10.1016/j.clon.2008.03.017

Original Article

Evaluation of a Method for Grading Late Photographic Change in Breast Appearance after Radiotherapy for Early Breast Cancer

J. S. Haviland*, A. Ashton†, B. Broad†, L. Gothard‡, J. R. Owen†, D. Tait‡, M. A. Sydenham*, J. R. Yarnold‡, J. M. Bliss*

*Clinical Trials and Statistics Unit (ICR-CTSU), Section of Clinical Trials, The Institute of Cancer Research, Sutton, Surrey, UK; †Gloucestershire Oncology Centre, Cheltenham, UK; ‡Department of Radiotherapy, Royal Marsden NHS Foundation Trust, Sutton, Surrey, UK

ABSTRACT:
Aims: **Serial photographs have been collected prospectively to evaluate the effect of radiotherapy on normal tissues in the breast. The aim of this study was to compare two methods of scoring radiation-induced changes.**
Materials and methods: **Five-year photographs of 400 patients randomised to receive either 42.9 or 39 Gy in 13 fractions to the whole breast after tumour excision of early breast cancer were compared with a post-surgery baseline and scored for change in breast appearance on a three-point graded scale. Two alternative methods of scoring using three observers were compared: (a) scores allocated independently, with independent resolution of discrepancies, and (b) scores allocated by consensus.**
Results: **Treatment effects estimated from the consensus and independent scores were very similar (odds ratio 1.89, 95% confidence interval 1.21–2.96 vs 2.28, 95% confidence interval 1.50–3.47, respectively). Agreement between the scores obtained from each method was reasonable, and the repeatability of the consensus method was good.**
Conclusions: **The consensus method of scoring photographic change in breast appearance seems to be no less sensitive to randomised dose as the independent method of assessment, but is much quicker to administer. The consensus method has been used to score over 3000 sets of photographs in the National Cancer Research Institute Standardisation of Breast Radiotherapy trial. Haviland, J. S. *et al*. (2008). *Clinical Oncology* 20, 497–501**

Key words: Breast cancer, normal tissue effects, photographic assessments, radiotherapy, scoring method

Introduction

The outcome of tumour excision and radiotherapy on the preserved breast in women with early breast cancer has been recorded using a graded cosmetic scale (excellent, good, fair, poor) applied by the treating clinician [1–4]. Cosmesis as an aesthetic judgement is certainly relevant, but is only valid when scored by patients than by external observers [5,6]. It is a valuable global outcome for judging the overall success or failure of surgery and radiotherapy, but it is not the most sensitive end point of radiation adverse effects [7]. This is because radiation effects are not the only factors influencing the breast cosmesis score, as surgical deficit, scar appearance and the woman's expectations are also important.

Serial photographs scored by external observers using objective criteria in comparison with a post-surgical baseline image have been used to score radiotherapy adverse effects for many years [8–13]. The precise methodology has varied, with some investigators projecting a calibrated grid on which objective linear measurements can be collected [8]. The post-surgical baseline photograph of the contralateral breast is needed to control for time-dependent changes unrelated to radiotherapy, including weight gain and effects of ageing. After allowance for these and surgical deficit in the treated breast, any additional changes in the treated breast are attributed to radiotherapy.

A change in breast appearance scored from photographs was the primary end point for the Royal Marsden Hospital/Gloucestershire Oncology Centre (RMH/GOC) trial of alternative fractionation regimens in women treated by breast conservation surgery and radiotherapy for early breast cancer [14]. The results of this randomised trial confirmed that changes in breast appearance scored from annual photographs compared with post-surgical baseline photographs provide a measure of late radiation change that is sensitive to a 10% difference in radiotherapy fraction size. In this trial, dedicated evaluation sessions were held, which required three observers to each independently score follow-up photographs. Discrepancies between observers

0936-6555/08/200497+05 $35.00/0

were re-scored independently. This method of scoring photographic change in breast appearance was assessed for reliability and repeatability, and scored highly using the kappa statistic. All results relating to the RMH/GOC trial have been published using these data [14].

A change in breast appearance, as assessed by photographs, was an important secondary end point in the National Cancer Research Institute (NCRI) Standardisation of Breast Radiotherapy (START) trial, which tested alternative fractionation regimens in 4451 women with early breast cancer [15]. In the START trial, more than 2000 women were recruited after breast-preserving surgery into a photographic sub-study, in which photographs were taken before radiotherapy and at 2 and 5 years. At a median follow-up of 5 years overall, the challenge of scoring more than 3000 pairs of photographs using the methodology used in the previous trial raised logistical difficulties. The original method was very time-consuming, due to independent re-scoring of discrepancies followed by a discussion to reach consensus if re-scoring did not achieve concordance. For this reason, an alternative non-independent scoring system was proposed, whereby three observers conferred and agreed a score by consensus, without making prior independent assessments. Apart from routine monitoring of reproducibility by test–retesting of a random sample, each pair of photographs was assessed only once, with considerable time saving. In order to test this simplified method, a subset of the RMH/GOC fractionation trial photographs was re-scored using the consensus method, to test if the treatment effect found using the independent scoring method could be reproduced. If so, the consensus method would be used for assessments of change in breast appearance from photographs in the NCRI START trial.

Materials and Methods

Details of the RMH/GOC trial have been published elsewhere [14]. Briefly, between 1986 and 1998, 1410 patients were enrolled and randomised to receive one of three radiotherapy regimens after local tumour excision of early stage breast cancer. The control schedule was 50 Gy in 25 fractions, and the two test schedules were 39 or 42.9 Gy in 13 fractions of 3.0 or 3.3 Gy, all delivered over 5 weeks. The primary end point was change in breast appearance, obtained from a comparison of photographs taken after surgery and during follow-up.

Photographic Assessments of Change in Breast Appearance

Frontal photographs of both breasts were taken under standard conditions after primary surgery, before radiotherapy, and repeated annually for 5 years and then at 10 years. Two photographs were taken of the patient's trunk region, one with the hands resting on the hips, the other with the arms raised above the head. Follow-up photographs were terminated in the case of local relapse, further breast surgery, declining health or patient refusal.

Photographs were scored under standardised conditions by three observers blind to patient identity, fractionation allocation and year of follow-up. Comparisons were always based on photographs at two time points, one showing postoperative appearance and the other showing breast appearance at a later time point. Changes in the contralateral breast made it possible to distinguish radiotherapy effects from other time-related changes. A change in breast appearance compared with the postoperative baseline was scored on a three-point graded scale (none/minimal = 0, mild = 1, marked = 2) based on a change in breast size, mainly shrinkage and distortion (Fig. 1). Each scoring session began with a training set of photographs to ensure standardisation of scoring criteria between observers.

Independent Scoring Method

Each of the three observers (two oncologists [JRO and JY] and one nurse specialist [AA/BB]) independently allocated a score (three-point scale described above) for a change in breast appearance. Where there was a discrepancy, the photographs were again scored independently, together with an additional random subset (about 10% of the whole series) to investigate repeatability. No information on original scores was provided when discrepancies and the 10% random subset were re-scored. When the discrepancy persisted, the modal category was chosen and in cases where no modal category was obvious, observers were asked to reach a consensus by open discussion.

Consensus Scoring Method

A score for a change in breast appearance was allocated by consensus between three conferring observers (two radiation oncologists [DT and JY], and one trial co-ordinator [LG]). Taking it in turns, one observer offered a grading that the other two observers either accepted or challenged. If challenged, discussion followed until consensus was reached. The same three-point graded scale was used as in the independent method. A random 10% subset was re-scored on the same day as the initial consensus scores in order to test for repeatability.

Sample Size

In the RMH/GOC trial, the main treatment comparison of interest was between the two experimental fractionation schedules (i.e. 39 and 42.9 Gy in 13 fractions). For the purposes of verifying the treatment effect, it was not considered informative to re-score photographs in women allocated 50 Gy. The largest prevalence of change in breast appearance in the RMH/GOC trial was at 5 years, and the treatment effect obtained at 5 years was similar to that obtained from survival analysis across all photographic assessments. Hence, the re-scoring focussed on year 5 photographs. From the previous data obtained using the independent scoring method, rates of change (mild and marked) in breast appearance at 5 years were 35.5% (95%

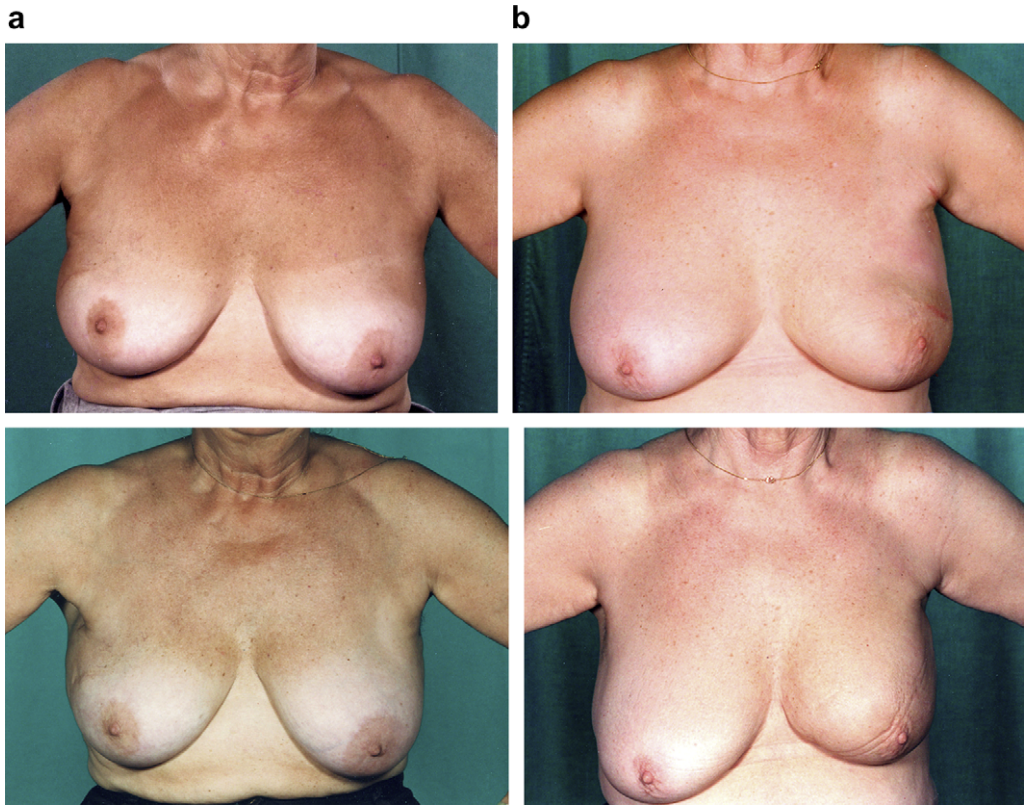

Fig. 1 — Serial photographs taken after surgery and before radiotherapy (top) and during follow-up (bottom) illustrating: (a) no change in breast appearance (60 months post-RT right breast) and (b) a marked change in breast appearance (30 months post-RT left breast).

confidence interval 29.8—41.7) in the 42.9 Gy group and 19.4% (95% confidence interval 14.9—24.9) for 39 Gy (odds ratio 2.28, 95% confidence interval 1.50—3.47). Two hundred patients per group (total 400) would enable the treatment effect to be estimated to within a similar degree of precision (95% confidence interval for odds ratio 1.45—3.60).

Because the consensus scoring was carried out at RMH and the photographs were stored locally at each of the centres (RMH and GOC), most of the sample was taken from RMH patients, for convenience. All RMH patients treated with 42.9 or 39 Gy and with baseline and 5-year photographs available were included. The remainder of the sample was randomly selected from eligible GOC patients.

Statistical Methods

Given the limitations of the available sample size, the aim of this analysis was not to test for a significant difference between the treatments within the RMH/GOC trial, but to compare the estimates of treatment effect obtained using the two scoring methods. The effect of fractionation schedule on the change in breast appearance at 5 years was estimated by calculating an odds ratio together with a 95% confidence interval. Pairs of scores obtained from the consensus and independent methods were compared within patients and the weighted kappa statistic calculated as a measure of agreement. The test of symmetry [16] was used to test whether one method produced higher scores for change in breast appearance compared with the other method of assessment.

Results

Four hundred year 5 photographs were scored by three observers reaching consensus by open discussion for a change in breast appearance compared with the post-surgical baseline. The pairwise scores using the independent and consensus methods are shown in Table 1. The totals in the table show that mild and marked changes in breast appearance were scored in 88 (22.0%) and 16 (4.0%) patients using the independent method compared with 97 (24.2%) and 32 (8.0%) using the consensus method, respectively. Any change in breast appearance (mild or marked) at 5 years was recorded using the consensus scores in 77 (39.3%, 95% confidence interval 32.5—46.5) and 52 (25.5%, 95% confidence interval 19.8—32.1) patients randomised to 42.9 and 39 Gy in 13 fractions, respectively. The odds ratio for change in breast appearance at 5 years for 42.9 Gy vs 39 Gy was 1.89 (95% confidence interval 1.21—2.96), which compared favourably with the treatment effect estimated using independent observers (odds ratio 2.28, 95% confidence interval 1.50—3.47).

A comparison of the consensus scores with independent scores showed an observed agreement of 77.2%, with

Table 1 — Comparison of scores for change in breast appearance at 5 years after radiotherapy obtained using independent and consensus methods: 400 patients randomised into a breast radiotherapy fractionation trial

Consensus score	Independent score			Total for consensus score (column %)
	None	Mild	Marked	
None	248	23	0	271 (67.7)
Mild	46	48	3	97 (24.2)
Marked	2	17	13	32 (8.0)
Total for independent score (row %)	296 (74.0)	88 (22.0)	16 (4.0)	400 (100)

a weighted kappa statistic of 0.56, indicating 'moderate agreement'. Similar levels of agreement were obtained for each of the fractionation schedules, with weighted kappa statistics of 0.56 and 0.53 for 42.9 and 39 Gy, respectively. Overall, it seems that a change in breast appearance (mild or marked) was scored more frequently using the consensus method, although numbers of marked events in particular were small. From Table 1 it can be seen that 26 pairs of photographs (6.5% of 400) were scored less severely using the consensus method compared with the independent method (i.e. numbers above the diagonal in the table) and 65 (16.2%) more severely (i.e. numbers below the diagonal). The test of symmetry was highly significant ($\chi^2 = 19.5$, df $= 3$, $P < 0.001$), indicating that the distributions of scores in the table were not symmetric, with higher scores for a change in breast appearance using the consensus method. All except two of the discrepancies between the methods differed by only one category. For these two patients, their 5-year scores were checked against scores allocated for years 1—4, and the consensus score was found to be more consistent with previous years than the independent score. A random 10% subset of RMH patient photographs was re-scored in order to test for repeatability of the consensus method. This showed good agreement (86.2%), with a weighted kappa statistic of 0.77.

Discussion

Assuming an α/β value of 3 Gy for a late change in breast appearance, 39 and 42.9 Gy in 13 fractions over 5 weeks are equivalent to 46.8 and 54 Gy, respectively, in 2.0 Gy equivalents. The odds ratio of about 2 for a change in breast appearance over this range of dose intensity is a vivid reminder of the steepness of the dose—response curve for late adverse effects, representing a γ-value of 1.8 [14]. The magnitude of treatment effect obtained using the consensus method of scoring a change in breast appearance from photographs was slightly less than that based on the original independent scores. When three observers conferred, the odds ratio for a radiation-induced change in breast appearance after 42.9 Gy vs 39 Gy in 13 fractions was 1.89 compared with 2.28 when observers scored independently. These ratios of effect fall within the same range (95% confidence interval 1.21—2.96 vs 1.50—3.47, respectively), suggesting that the consensus method of scoring photographic change is no less sensitive to randomised dose than the independent method. Although there are possible explanations for differences in scores between the two methods, these are unlikely to affect the odds ratio estimate of treatment effect as all assessments were carried out blind to radiotherapy schedule.

Agreement between the sets of scores obtained using the two methods was reasonable, and the consensus method showed a high level of repeatability between repeat scores. However, there was evidence of photographs being allocated higher scores for a change in breast appearance using the consensus method. The scoring of independent and consensus sets was separated by more than 3 years, and it is possible that despite training sessions to ensure that the same criteria were applied, some differences nevertheless remained. It is possible that thresholds for scoring late normal tissue effects have changed over the years, as fewer and milder effects are generally seen nowadays due to improvements in radiotherapy planning and delivery. Therefore, this might lead to changes in breast appearance being assessed as more marked using current expectations, compared with a few years ago. It should be noted that only one (JY) of the three observers contributed to both independent and consensus scores. In the consensus scoring, a female (DT) replaced a male (JRO) radiation oncologist and a female trial co-ordinator (LG) replaced a female nurse oncologist (AA/BB). However, a comparison of JY's original independent scores with the consensus scores showed a similar level of agreement (weighted kappa $= 0.54$) as for the comparison with the final score obtained from all three independent observers (weighted kappa $= 0.56$). This indicates that any differences between the independent and consensus scores are unlikely to be due to a change in observers.

Analysis of the time taken to score the photographs confirmed that the consensus method is quicker, allowing an average of over 150 assessments per hour compared with less than 100 per hour using the independent method. This difference underestimates the savings in time, as, using the independent scoring method, any discrepancies require photographs to be recalled for re-scoring, and recalled once again if the scores are still discrepant at the second review. The independent method also uses a computer program in which the three sets of scores are entered in batches during the sessions, and which generates lists of discrepancies to be re-scored, thereby adding to the administrative complexity. Data from the original independent scoring sessions for all follow-up photographs from the trial show that the system of re-scoring discrepancies leads to an increase in the total number of assessments required per photograph of 40% overall.

Whichever scoring method is adopted in the current context, it is important to start with a training set of photographs that show the full range of treatment effects and to agree criteria for grading change. The first factor to allow for is the surgical deficit, using the contralateral

breast in the baseline photograph as a control. Subsequent weight change or age-related changes are also controlled by reference to the contralateral breast. As mentioned in the Materials and Methods section, the most obvious and frequent change is breast shrinkage, presenting as a loss of volume affecting the whole organ. A change in post-surgical breast shape (distortion) with time may be seen in addition to a loss of volume, a reliable indicator of underlying fibrosis. Either way, one of the key indicators of volume loss and/or distortion is nipple displacement. A shift in the position of the nipple has been successfully used as the basis for scoring treatment outcome in other published photographic scoring systems [8,10–12,17]. Linear measurements of nipple displacements can be collected from photographs in an objective and reproducible manner. However, the current analysis suggests that a subjective assessment of change can also be applied very quickly in a reproducible manner to detect small differences in radiotherapy dose.

Although telangiectasia is easily seen on exposed surfaces of the breast, it is usually confined to the inframammary fold of heavy-breasted women, where it reflects loss of skin sparing. Telangiectasia was, therefore, not considered in scoring the photographic phenotype. In patients with a marked change in breast appearance, the skin frequently looks shiny, a difference interpreted as evidence of skin dryness and atrophy. Although not formally included in the scored phenotype, this change in appearance is a common feature in women with marked breast shrinkage and/or distortion. Other important elements of the late radiation effects that cannot be scored from photographs include breast induration (assumed to represent fibrosis) and underlying damage to muscle, ribs, lungs or heart. These require additional assessments, including simple palpation.

In conclusion, the logistics of the consensus method far outweigh the original independent method in terms of time taken and ease of recording scores. Since this validation study was carried out, a team of three (JRO, JY, LG) has scored change in breast appearance from the photographs taken in the NCRI START trial using the consensus method. It is proposed that the new method will be used for all photographic assessments of change in breast appearance in future radiotherapy trials.

Acknowledgments. We thank all the patients who participated in the trial. Thanks are also due to a number of colleagues at RMH/GOC/ICR who have been involved with the scoring sessions and data management, especially Janis Homewood. The trial was supported in part by Marks and Spencer plc. The ICR-CTSU receives funding from Cancer Research UK.

Author for correspondence: J. S. Haviland, Clinical Trials and Statistics Unit (ICR-CTSU), Section of Clinical Trials, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. Tel: +44-208-722-4042; Fax: +44-208-770-7876; E-mail: jo.haviland@icr.ac.uk

Received 19 November 2007; received in revised form 13 March 2008; accepted 31 March 2008

References

1 Clarke D, Martinez A, Cox RS. Analysis of cosmetic results and complications in patients with stage I and II breast cancer treated by biopsy and irradiation. *Int J Radiat Oncol Biol Phys* 1983;9(12):1807–1813.
2 Ray GR, Fish VJ. Biopsy and definitive radiation therapy in stage I and II adenocarcinoma of the female breast: analysis of cosmesis and the role of electron beam supplementation. *Int J Radiat Oncol Biol Phys* 1983;9(6):813–818.
3 Olivotto IA, Rose MA, Osteen RT, *et al*. Late cosmetic outcome after conservative surgery and radiotherapy: analysis of causes of cosmetic failure. *Int J Radiat Oncol Biol Phys* 1989;17(4):747–753.
4 Kurtz JM. Impact of radiotherapy on breast cosmesis. *Breast* 1995;4:163–169.
5 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):153–159.
6 Sneeuw KC, Aaronson NK, Yarnold JR. Cosmetic and functional outcomes of breast conserving treatment for early stage breast cancer. 2. Relationship with psychosocial functioning. *Radiother Oncol* 1992;25(3):160–166.
7 Pezner RD, Lipsett JA, Vora NL, *et al*. 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):1117–1119.
8 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):159–167.
9 Van Limbergen E, Rijnders A, van der Schueren E, *et al*. Cosmetic evaluation of breast conserving treatment for mammary cancer. 2. A quantitative analysis of the influence of radiation dose, fractionation schedules and surgical treatment techniques on cosmetic results. *Radiother Oncol* 1989;16(4):253–267.
10 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):667–676.
11 Christie DRH, O'Brien MY, Christie JA, *et al*. A comparison of methods of cosmetic assessment in breast conservation treatment. *Breast* 1996;5:358–367.
12 Christie D, Sharpley C, Curtis T. Improving the accuracy of a photographic assessment system for breast cosmesis. *Clin Oncol* 2005;17(1):27–31.
13 Yarnold J. Assessing breast cosmesis after radiotherapy: what do we want to measure. *Clin Oncol* 2005;17:25–26.
14 Yarnold J, Ashton A, Bliss J, *et al*. Fractionation sensitivity and dose response of late adverse effects in the breast after radiotherapy for early breast cancer: long-term results of a randomised trial. *Radiother Oncol* 2005;75(1):9–17.
15 Venables K, Miles EA, Hoskin PJ, *et al*. Verification films: a study of the daily and weekly reproducibility of breast patient set-up in the START trial. *Clin Oncol* 2005;17(5):337–342.
16 Bowker AH. A test for symmetry in contingency tables. *J Am Stat Assoc* 1948;43(244):572–574.
17 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):575–578.