Evaluating Intra and Inter-observer bias in the Cosmetic rating for random vs. serial assessment of Breast photographs

Preeti Belani  
Tata Memorial Centre, Homi Bhabha National Institute

Rima Pathak  
Tata Memorial Centre, Homi Bhabha National Institute

Shraddha Kenekar  
Tata Memorial Centre, Homi Bhabha National Institute

Gaurika Pokale  
Tata Memorial Centre, Homi Bhabha National Institute

Pallavi Rane  
Tata Memorial Centre, Homi Bhabha National Institute

Ashwini Chalke  
Tata Memorial Centre [TMC]

Tabassum Wadasadawala  
twadasadawala@actrec.gov.in

Tata Memorial Centre, Homi Bhabha National Institute

Article

Keywords: SAS [Series assessment setting], RAS [Random assessment setting], RO [Radiation Oncologist], RN [Registered nurses] BCCT [Breast Cancer Conservation treatment], BCS [Breast conserving surgery]

Posted Date: October 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3420202/v1

License: 😊 This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Abstract

BACKGROUND

This study was done to assess inter and intra-rater bias in rating of cosmesis, when breast photographs were evaluated serially or randomly by a panel of six members having varying years of experience.

METHODS

Cosmetic assessment was done subjectively for 175 images [of 50 unilateral breast cancer patients for whom at least 3 images were collected], that were arranged serially from baseline to follow up in chronological order termed ‘serial assessment setting’ [SAS]. For ‘random assessment setting’ [RAS], all images was randomly arranged for assessment. Objectively assessment was also done using BCCT.core. Kappa index was calculated for agreement between the RAS and SAS rating for the 3 panellists’ groups and with BCCT.core.

RESULTS

Good agreement [kappa 0.659] was found between the mean panel cosmetic scores for both SAS and RAS. Fair agreement was found when subjective RAS [k=0.301] and SAS [k=0.343] scores were compared with the BCCT.core, which was highest for the most experienced panellists with SAS k=0.387 and RAS k=0.436. Both SAS and RAS had good intra-rater reliability.

CONCLUSIONS

SAS improves the agreement with BCCT.core rating and may be used if validated in a larger cohort. The clinical experience of the panellist impacts cosmetic rating and must be considered before forming a panel.

INTRODUCTION

The impact of Breast Conserving therapy [BCT] has been monumental on quality of life. Assessment of Cosmesis and Quality of life forms an important endpoint for clinical research. Various studies have demonstrated high correlation between poor cosmetic outcome and low scores on quality of life indicators. [1, 2]

Subjective or objective cosmetic evaluation is integral to clinical trials related to breast conserving therapy. Subjective assessment may be done utilizing the Harris scale [3] The Harris scale reports the outcome on a 4 point-Likert scale [excellent, good, fair and poor] based on the degree of asymmetry between the treated and untreated breasts. Even though the Harris scale is the simplest method used to subjectively score cosmesis, it is prone to high variability due to various factors like the professional background of the rater, method of capturing and processing the image, etc.
The BCCT.core is a commonly employed software for objective evaluation that processes images and rates the cosmesis on a 4-point scale [Excellent, Good, Fair or poor] using a semi-automatic software utilizing predetermined points placed on a frontal image. It is developed and owned by Cardoso J, Cardoso M and INESC Porto Breast Research Group. [4]

Breast Reconstruction Assessment [BRA] is also an objective method of cosmetic evaluation. [5] In fact, all of these assessment scales come with inherent inconsistency and bias. Pezner et al have demonstrated that it is difficult to obtain consensus of cosmetic outcome with the commonly employed scales. [6]

Cosmetic rating on breast photographs is commonly done in clinical trials, mostly retrospectively by a panel of experts like in the UK START Trial B. [7] If the rating is done using breast photographs, factors affecting the image quality will also affect the cosmetic rating. Use of standardized methods to acquire the photographs under uniform room lighting conditions, distance of the camera from the patient, and specifications of the camera can minimize this variability. The rating done by a clinician is also often different from the rating done by patients. [8] It also varies with the experience of the rater [9], and fatigue of the rater [when a bulk of images have to be rated].

It is also possible that if a clinician rates the images in serial order chronologically, one patient at a time, the rating of consequent images might change after rating the first image, either negatively or positively, thereby creating a bias. It is noteworthy that routine cosmetic evaluation done in high volume clinics is random since the clinician is unlikely to compare the current cosmetic status with a previously taken image. Hence, when cosmetic analysis is done retrospectively, and images of the patient taken at different time points is available to influence the decision of the rater, the resultant rating might be different from actual assessments done in clinics [which are random]. At the same time, serial images may also guide the clinician to give a better informed global cosmetic rating to ultimately know the temporal changes [worsening or improvement] in the treated breast.

In an IEC approved study being conducted in our centre [CTRI/2020/01/022871], images of breast cancer patients who underwent breast conserving surgery [BCS] and adjuvant RT have been collected at baseline and follow up visits, for cosmetic evaluation by a panel. These images were utilized in the current study with the objective of evaluating inter and intra-rater bias for assessment of cosmesis in the serial vs. random assessment settings, and if the experience of the rater impacts the global cosmetic rating.

**MATERIALS AND METHODS**

**Study population and photographs:**

The study population consisted of 50 women diagnosed with unilateral breast cancer treated with Breast conserving surgery and adjuvant Radiation Therapy at Tata memorial Centre Mumbai, who had at least 3
photographs [baseline and any 2 follow up visits], taken at least 6 months apart. The demographic details are given in table no. 1.

Table no. 1: Demographic data of the study population [n = 50]

<table>
<thead>
<tr>
<th>Laterality</th>
<th>Age</th>
<th>Type of BCS</th>
<th>Axillary Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right: 26</td>
<td>&lt; 40 years: 15</td>
<td>Open Cavity: 33</td>
<td>Axillary sampling: 11</td>
</tr>
<tr>
<td>Left: 24</td>
<td>41–50 years: 22</td>
<td>Type I Oncoplasty: 13</td>
<td>Axillary clearance [I-II]: 8</td>
</tr>
<tr>
<td></td>
<td>51–60 years: 8</td>
<td>Type II Oncoplasty [LD flap]: 2</td>
<td>Axillary clearance [I-III]:31</td>
</tr>
<tr>
<td></td>
<td>61–70 years: 5</td>
<td>Type II Oncoplasty [LICAP flap]: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type III Oncoplasty [Reduction]: 1</td>
<td></td>
</tr>
</tbody>
</table>

BCS: Breast Conserving Surgery, LD: Latissimus Dosri, LICAP: Lateral Intercostal Artery perforator

Frontal breast photographs in sitting position, light background, hands on hip, excluding the face were uniformly captured using a 13 megapixel camera on the trial device by the trial coordinator. No professional photographer was hired. The images were incorporated in Google forms for rating by the panel.

**Sample size:**

Cardoso et al found a kappa value of 0.59 between the rating given by experienced panellists using the Harris scale, and the BCCT.core software. The sample size was 55 patients [1 image for each patient]. [9] Similarly, Heil et al found the inter-rater agreement as 0.41 [mk] when the same panellists evaluated the same images 1 hour apart, for the images taken immediately after surgery. The sample size was 50 patients [2 images for each patient, just after surgery and 1 year after surgery]. [10]

In a test for agreement between two raters using the Kappa statistic, a sample size of 174 subjects achieves 81% power to detect a true Kappa value of 0.55 based on a significance level of 0.05. Therefore, in the current study, 175 photographs were taken for assessment.

**Panel:**

The assessment was done by a total of 6 panellists, 2 Senior Breast RO [experience of over 10 years in Breast Oncology], 2 Junior RO [less than 2 years of experience in Breast Oncology], and 2 registered nurses [RN]. Thus three groups of panellists were created with varied levels of experience.

**Evaluation:**

Subjective Assessment: This was performed using Harris scale [shown in Table no. 2] by the 6 panellists.

Table no. 2: Harris scale for Cosmetic Rating
<table>
<thead>
<tr>
<th>Rating</th>
<th>Score</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>1</td>
<td>Treated breast nearly identical to untreated breast</td>
</tr>
<tr>
<td>Good</td>
<td>2</td>
<td>Treated breast slightly different from untreated breast</td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
<td>Treated breast clearly different from untreated breast but not seriously distorted</td>
</tr>
<tr>
<td>Poor</td>
<td>4</td>
<td>Treated breast seriously distorted</td>
</tr>
</tbody>
</table>

The assessment was done using Google forms in which the Harris scale was populated and breast photographs were incorporated as per the study identification number.

When the photographs of each patient were arranged serially from baseline to follow up, it was termed as ‘serial assessment setting’ [SAS] whereas in the ‘random assessment setting’ [RAS], all images were randomly arranged. Both SAS and RAS were done by the six panellists independently using the 4 point Harris scale.

To avoid fatigue, the images were rated in 2 sets of 25 patients each [SAS followed by RAS]. A washout time of one week was given between SAS and RAS.

Objective Assessment: Cosmetic rating was done using BCCT.core software for each image by one of the panellists and the global rating noted. Various other indices were also produced by the software which are beyond the utility of this study and hence not reported.

**Statistical Analysis:**

- The assessments acquired using subjective and objective methods were entered into IBM SPSS Version 24.
- Mean Panel score was calculated for all images for SAS and RAS. Fractions were rounded off to nearest integer.
- Cohen Kappa [shown in table no. 3] was used to determine the agreement between following [A p value of 0.05 was considered statistically significant]:
  - Mean SAS and Mean RAS
  - Mean SAS with BCCT.core
  - Mean RAS with BCCT.core
  - Mean SAS and RAS for 3 panellists groups with BCCT.core

Table no. 3: Kappa values
\[
\begin{array}{|c|c|}
\hline
\text{Kappa index} & \text{Agreement} \\
\hline
< 0.2 & \text{Poor} \\
0.2–0.4 & \text{Fair} \\
0.4–0.6 & \text{Moderate} \\
0.6–0.8 & \text{Good} \\
0.8–1.0 & \text{Very good} \\
\hline
\end{array}
\]

- Cronbach's alpha coefficient was calculated separately for SAS and RAS to understand the inter-rater reliability between different raters. A value of 0.7 or higher was considered reliable.

**RESULTS**

In our study, we found that the agreement between mean RAS and mean SAS scores was good [0.659 kappa]. For both SAS and RAS, maximum images were rated as ‘good’, followed by ‘fair’ and then ‘excellent’. Only a small no of images [12 in SAS and 11 in RAS] were rated as ‘poor’ by the panel [Figure no. 1]. The good agreement demonstrates the acceptability of either methods for assessment.

The measure of agreement [kappa] of subjective methods with BCCT.core [shown in Figure no. 2] was 0.343 between SAS and BCCT.core, and 0.301 between RAS and BCCT.core. Both these values translate to ‘Fair’ agreement, however, slightly higher value of SAS with BCCT.core demonstrates the potential utility of providing Serial images during panel assessment.

The agreement between rating of the baseline images and the 1\textsuperscript{st} follow up images was slightly different for SAS [0.342] and RAS [0.442] [Supplementary table 1]. Also, agreement with BCCT was higher for SAS [0.430] than RAS [0.255] the first follow up as compared to baseline, showing that the agreement with objective evaluation tends to improve with time [Supplementary table 2].

The inter-rater reliability for both SAS and RAS was found to be 0.888 [Cronbach's alpha], which demonstrates high reliability of both methods [Figure no. 3].

In terms of experience, the panellists with highest experience [Breast RO with more than 10 years of experience] had the highest agreement with BCCT.core in both SAS and RAS [RAS slightly higher than SAS], as compared to the other two groups of panellists with lesser experience [Figure no. 4]. The trendline for SAS and RAS show a downward trend as the experience of the rater decreases.

The agreement also varied with the type of surgery [Table no. 4]. In our study, 33 patients had undergone open cavity BCS procedures in whom we found good agreement between SAS and RAS [0.733] as compared to moderate agreement in patients having undergone Oncoplasty [0.497]. For Oncoplastic procedures, agreement with BCCT.core [objective method] improved from fair to moderate when serial images were provided to the panel.
Table no. 4: Variation in agreement with the type of Surgery

<table>
<thead>
<tr>
<th>Type</th>
<th>1st Method of assessment</th>
<th>2nd Method of assessment</th>
<th>Kappa value</th>
<th>Interpretation</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Cavity</td>
<td>BCCT.core</td>
<td>SAS</td>
<td>0.300</td>
<td>Fair</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>BCCT.core</td>
<td>RAS</td>
<td>0.299</td>
<td>Fair</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>SAS</td>
<td>RAS</td>
<td>0.733</td>
<td>Good</td>
<td>0.00</td>
</tr>
<tr>
<td>Oncoplasty</td>
<td>BCCT.core</td>
<td>SAS</td>
<td>0.408</td>
<td>Moderate</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>BCCT.core</td>
<td>RAS</td>
<td>0.288</td>
<td>Fair</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>SAS</td>
<td>RAS</td>
<td>0.497</td>
<td>Moderate</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**DISCUSSION**

This study intended to evaluate the presence of bias in SAS and RAS [intra-observer], along with difference in the inter-observer rating. We found good agreement between SAS and RAS, with SAS having slightly better agreement with BCCT.core, suggesting that subjective rating is closer to objective rating [BCCT.core] if breast photographs are evaluated serially. This parameter has never been evaluated in historical studies of similar nature, and hence may be considered for a detailed analysis. The study also establishes the influence of the experience of the rater and the composition of panel for cosmetic assessment.

The cosmetic assessment of the treated breast is often subjectively done during routine follow up in the clinics, as performing a BCCT.core analysis is not feasible in high volume centres since it requires acquiring breast photograph and it’s processing in a dedicated software. Using BCCT.core is a time-consuming process that limits its utility. On the other hand, BCCT.core has become very popular for objective assessment in research setting. Heil et al have found very good agreement between different users of the semi-automated BCCT.core software. [10] Stored photographs can hence be scored objectively, as well as subjectively by an individual physician or by a panel of experts retrospectively. Panel assessment is a standard approach adopted in large breast cancer trials. [11]

In our study, we found that the mean scores for SAS and RAS were in good agreement [kappa 0.659]. This means that either of these can be adopted for cosmetic assessment when a panel has to evaluate several images. A fair agreement of both SAS and RAS was noticed with BCCT.core, though the value was slightly higher for SAS [0.343] than RAS [0.301].This indicates that providing serial images of the earlier visits is likely to improve the rating of the current image [that has to be rated]. However, this finding needs to be validated in a larger cohort before implementation.

No prior studies have evaluated the significance of providing serial images [SAS] however, in a study conducted by Heil et al, images were evaluated twice by a panel, one hour apart, taken immediately after
and 1 year after surgery. They found that there was significant intra-rater agreement between both the ratings. The kappa values were 0.41 shortly after surgery and 0.32 one year after surgery, demonstrating either worsening of agreement as more time passed after treatment or lack of availability of serial images as they had only 1 image for comparison. [10] In our study, we found that intra-rater agreement for SAS between baseline and 1st follow up image was lower [0.342] as compared to RAS [0.442].

The duration of time between the treatment and the rating is also significant. This may be due to the natural healing process of the tissue, or due to a timeline bias where a freshly treated tissue appears better or worse than it actually is. In our study, we found better agreement with BCCT.core for both SAS and RAS at 1st follow up as compared to the baseline [Supplementary Table 2], contrary to Heil et al. [10]

Preuss et al found fair agreement of 0.37 [Kappa] between median subjective scores [panel] and BCCT.core. Agreement was higher with LD flap [kappa 0.51] and delayed reconstruction [0.60] as compared to TRAM flap reconstruction and immediate reconstruction. [12] In our study, 33 patients had undergone open cavity BCS, and only 17 patients had undergone Oncoplastic procedures and we found good agreement between SAS and RAS [0.733] for open cavity procedures, and moderate agreement [0.497] for Oncoplastic procedures. For Oncoplastic procedures, the agreement with BCCT.core improved from fair to moderate when serial images were provided for assessment, again indicating the positive influence of serial images. This observation also needs to be tested in a larger cohort.

Heil et al also found only slight to fair inter-rater agreement within each group [5 groups, each having 5-7 members of uniform profile, including breast surgeons, nursing staff, patients, medical and non-medical students], and the agreement was again found to be better immediately after surgery as compared to the agreement one year after surgery. [10] Corica et al have also shown modest inter-rater agreement between patients, doctors, nurses and BCCT.core in their study [13]. In our study, inter-rater kappa was not calculated as the three groups in the panel consisted of only 2 members each.

Haloua et al have shown that BCCT.core software was substantially in agreement with the mean overall panel score [panel consisting of 10 members] for cosmetic outcome [weighted kappa: 0.68, indicating substantial agreement]. This is higher than the agreement with BCCT.core in our study [SAS 0.343 and RAS 0.30]. The higher agreement in their study may be due to a larger panel in their study [10 member panel], quality of photography, or the presence of highly experienced clinicians in their panel in larger proportion. Their panel consisted of 2 experienced breast surgeons, 2 surgical residents, 4 experienced plastic surgeons and 2 lay persons. [14]

Yu et al determined the kappa statistic between the patients’ self-reported score and BCCT.core. They found low kappa value of 0.12 [indicating slight agreement]. [15]

The composition of a panel must be carefully chosen. A good panel has the potential of having high agreement with BCCT.core. For e.g., as discussed above, for Haloua et al, the panel consisted of ten members, out of which at least six were experienced surgeons [breast and plastic surgeons], and 2 surgical residents and only two lay persons. [14]. In the Cambridge Breast IMRT trial, the cosmetic
evaluation was done by a team of seven clinicians [four oncologists, one radiographer, one surgeon, and one breast care nurse] in a panel of any three at a time, and Mukesh et al concluded that improved dose homogeneity with simple IMRT translates into superior overall cosmesis. [16] In the EORTC ‘Boost v/s no boost’ trial, a panel cosmetic evaluation was performed by a panel consisting of five members. [17]. In the UK START B Trial, changes in the appearance of breast photographs was observed by three blinded observers. [7]. In line with literature, the current study also involved a panel comprising of six members.

Professional background and sex of the rater can also potentially affect the rating. In our study, all panellists were females and from Radiation Oncology background. No comprehensive study has evaluated the bias between Radiation and Surgical oncology background ratings, as most panels are an amalgamation of the two specialities. [14] However, such a study may be conducted, as a surgeon will have deeper understanding of surgical nuances, techniques and practices which may create a bias.

In our study, the panel consisted of three groups, two panellists in each group [two senior breast RO, two junior RO and two RN]. We found that the agreement with the BCCT.core improved with the increasing experience of the panellist. The senior Breast RO had the highest agreement with BCCT.core, suggesting that experience of the rating panellist is significant in cosmetic rating [Figure no. 4]. Similar results were seen in a study conducted by Cardoso et al. They found higher inter-observer agreement for experienced observers [multiple kappa 0.59] as compared to medium and inexperienced observers. Overall inter-observer agreement [for a total of 13 observers] was found to be 0.33 [multiple kappa]. They concluded that a homogeneous group of observers with experience in breast cancer conservative treatment will provide better inter-observer agreement than a mixed group involving clinicians with different levels of expertise.[9]

The strength of our study lies in the simple design, testing of raters with different experience levels, and being the first study to report impact of study setting on cosmetic rating. We included a mix of breast patients with different clinical stage, breast size, skin colour, and different surgeries to reflect a real-world clinical setting. Though simple in design, our study has limitations. Our study does not evaluate the influence of professional background [Plastic surgeon vs. Oncosurgeon vs. Radiation Oncologist] on the rating. Only frontal photographs were assessed objectively and subjectively, hence if the scar was not visible on the frontal photograph, it was not considered by the panel for cosmetic rating. All the patients did not have uniform number of photographs. Our photographs were not captured by a professional photographer and hence, minimal variations were observed in the quality of the images, the background, the distance, and the patient posture. We tried to overcome these limitations by post-processing the photographs for sharpness, quality and contrast and expect that the impact of these issues would be uniform on all the raters and on both SAS and RAS. The study did not intend to compare subjective and objective methods of assessment with each other.

CONCLUSION
There is low risk of bias when comparing RAS and SAS for breast cosmesis rating. SAS may improve the agreement with BCCT.core rating especially for oncoplastic procedures and may be used in clinical settings if validated in a larger cohort. Agreement with objective methods also varies with time after treatment. The clinical experience of the panellist significantly impacts cosmetic rating and must be considered before forming a panel for cosmetic assessment.

**Declarations**

**We declare no conflict of interest.**

**All study related procedures were carried out as per good clinical practice and ethical guidelines.**

**Informed consent was obtained from all participants regarding capturing the photographs and assessment of cosmesis under an IEC approved study at our institute (CTRI/2020/01/022871).**

**Data Availability:**

- The photographs taken for the assessment are available with the corresponding author. However they cannot be shared due to confidentiality and sensitive nature.
- The dataset utilized for statistical analysis is available with the corresponding author and data sharing is solely at the discretion of the corresponding author's institution policies.

**References**


**Figures**

![Figure 1](image)

**Figure 1**

**Series and Random assessment [SAS and RAS]**

River plot generated using Sankymatic.com
Figure 2

Measure of Agreement

Figure 3

Inter-rater Reliability
Inter-rater reliability for SAS and RAS

Figure 4

Agreement variation with experience of the Rater

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplements.pdf