EPOS™
Electronic Presentation Online System

European Society of Radiology

Observer variability in accept/reject classification of clinical image quality in mammography

Poster No.: C-2432

Congress: ECR 2015

Type: Scientific Exhibit

Authors: P. Whelehan[1], M. Hartswood[2], A. Mumby[3]; [1]DUNDEE/UK, [2]Edinburgh/UK, [3]Glasgow/UK

Keywords: Quality assurance, Observer performance, Mammography, Breast

DOI: 10.1594/ecr2015/C-2432

Aims and objectives

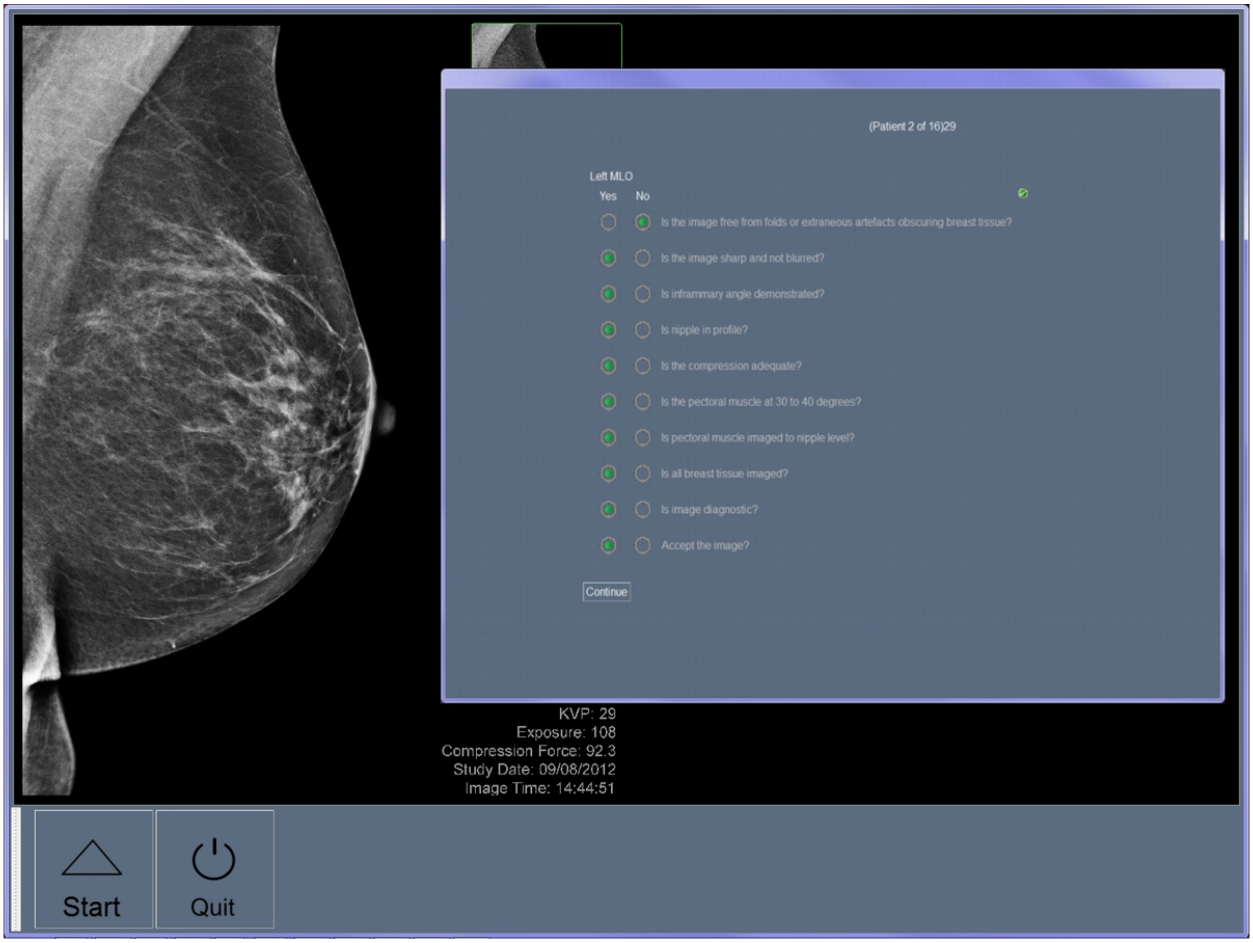
Image quality in mammography is widely accepted to be an important factor in the detection of breast cancer, and published evidence provides some support for this almost self-evident position[1,2]. Robust decision-making on when an image is of acceptable quality for interpretation is therefore vital.

In earlier work, we developed a computer-based training intervention (Figures 1-3), intended to improve mammography practitioners' decision-making with respect to whether an image should or should not be repeated. As well as its potential use in the standardisation, training and assessment of mammography image quality evaluation among clinical mammographers, the intention was to help mitigate the increase in unnecessary technical repeats which has been observed in breast screening with the transition to digital mammography.

The computerised training tool presents mammograms to the participants and collects their opinions on whether to accept or reject each image. Participants' judgements on whether the images meet specific detailed quality criteria are also recorded. The participants' decisions are compared to the inbuilt reference standard decisions and immediate feedback is provided. Participants' overall performance in evaluating image quality can be assessed against the reference standard.

This EPOS presentation describes the development of a reference standard for the training and test sets of mammography images to be used within this new computerised clinical image quality evaluation training tool. The study assesses the observer variability among the experts who contributed to the reference standard.

Images for this section:



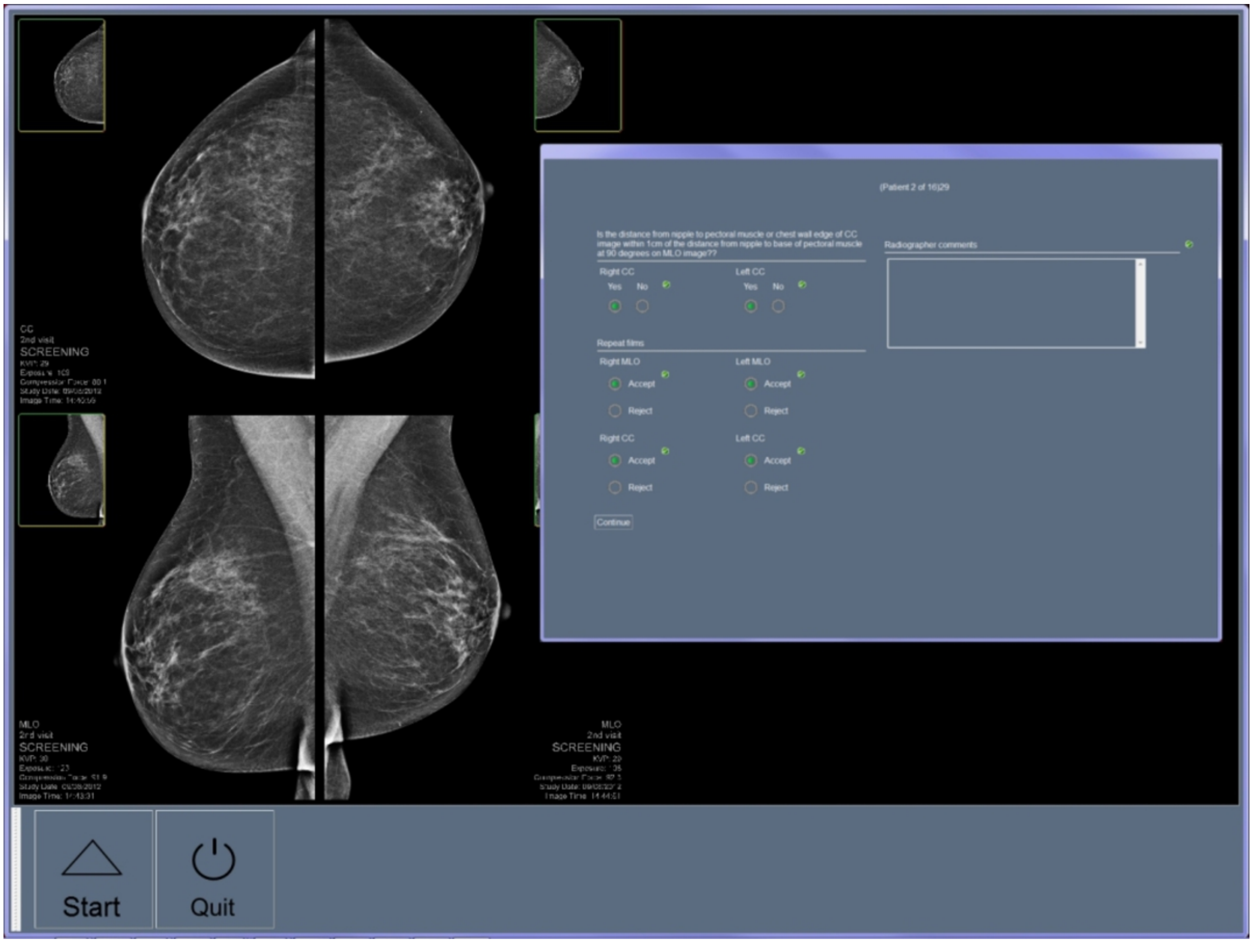
(Patient 2 of 16)29

Left MLO

Yes	No	
○	●	Is the image free from folds or extraneous artefacts obscuring breast tissue?
●	○	Is the image sharp and not blurred?
●	○	Is inframmary angle demonstrated?
●	○	Is nipple in profile?
●	○	Is the compression adequate?
●	○	Is the pectoral muscle at 30 to 40 degrees?
●	○	Is pectoral muscle imaged to nipple level?
●	○	Is all breast tissue imaged?
●	○	Is image diagnostic?
●	○	Accept the image?

KVP: 29
Exposure: 108
Compression Force: 92.3
Study Date: 09/08/2012
Image Time: 14:44:51

Fig. 1: A screen within the training tool, showing how participants record quality deficits per image and whether the image is acceptable



(Patient 2 of 16)29

Is the distance from nipple to pectoral muscle or chest wall edge of CC image within 1cm of the distance from nipple to base of pectoral muscle at 90 degrees on MLO image??

Right CC: ● Yes ○ No

Left CC: ● Yes ○ No

Repeat films

Right MLO: ● Accept ○ Reject

Left MLO: ● Accept ○ Reject

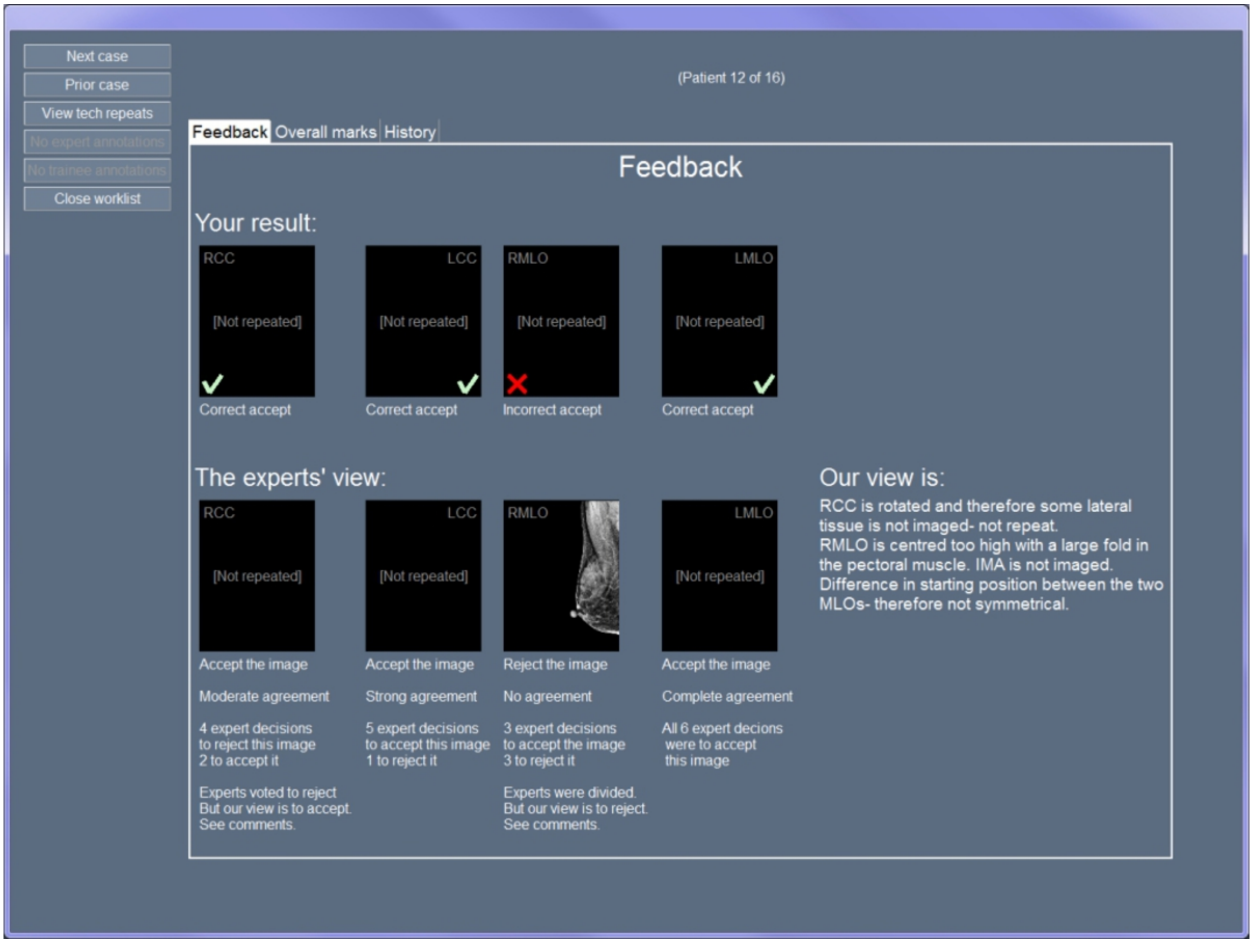
Right CC: ● Accept ○ Reject

Left CC: ● Accept ○ Reject

Radiographer comments

Continue

Fig. 2: A screen within the training tool, providing an overview of the examination and the opportunity to compare images and review acceptability decisions.



(Patient 12 of 16)

Feedback

Your result:

RCC	LCC	RMLO	LMLO
[Not repeated]	[Not repeated]	[Not repeated]	[Not repeated]
✓	✓	✗	✓
Correct accept	Correct accept	Incorrect accept	Correct accept

The experts' view:

RCC	LCC	RMLO	LMLO
[Not repeated]	[Not repeated]		[Not repeated]
Accept the image	Accept the image	Reject the image	Accept the image
Moderate agreement	Strong agreement	No agreement	Complete agreement
4 expert decisions to reject this image 2 to accept it	5 expert decisions to accept this image 1 to reject it	3 expert decisions to accept the image 3 to reject it	All 6 expert decions were to accept this image
Experts voted to reject But our view is to accept. See comments.		Experts were divided. But our view is to reject. See comments.	

Our view is:

RCC is rotated and therefore some lateral tissue is not imaged- not repeat.
RMLO is centred too high with a large fold in the pectoral muscle. IMA is not imaged.
Difference in starting position between the two MLOs- therefore not symmetrical.

Fig. 3: A screen within the training tool, showing feedback provided to participants, comparing the participant's opinion with the expert consensus opinion

Methods and materials

The highly experienced lead radiographer from one of the national mammography training centres in the UK selected a sample of digital mammograms from the archives of a regional population-based breast screening programme. The image sample was purposively compiled to include a range of quality deficits which can occur during mammography, and to incorporate some cases where the decision on whether the image was acceptable or not was considered potentially challenging.

Forty-two bilateral two-view mammograms were selected (168 images). These were then reviewed by three further national expert mammography radiographers within the UK, to reach an authoritative consensus on whether each image was acceptable or not. Of the three experts, two were currently and one formerly employed as mammography educators and competency assessors. Two of the three were also qualified to interpret mammograms - a recognised extended role for radiographers in the UK - and one of the three was a former regional breast screening quality assurance radiographer.

Each image was reviewed in accordance with UK National Health Service Breast Screening Programme guidelines[3] for assessing image quality (Table 1). Additionally, the "1 cm rule" was applied, for evaluating the amount of tissue included on the cranio-caudal image in relation to the matching medio-lateral oblique image. Although no grading system was employed, except for Accept/Reject, outside of this study the expert reviewers ordinarily use the "PGMI" (Perfect, Good, Moderate, Inadequate) system[4] in their training and assessment practice.

The three experts recorded their opinions using the training tool software, and each repeated the review a minimum of one month later. Thus, there were six expert decisions recorded for each image, and both intra- and inter-observer reliability could be assessed. The findings captured by the software tool were exported to Microsoft Excel and imported into IBM SPSS Statistics v.21 for analysis. Cohen's kappa statistic, which is intended to assess the level of agreement above that which would be expected by chance, was used to measure reliability of the image acceptability assessments within experts and between pairs of experts. Fleiss' kappa was used to assess reliability across all three experts.

Table 1: NHSBSP Criteria for assessing clinical image quality

Acceptability Criteria:

medio-lateral oblique

Whole breast imaged

Nipple in profile

Correct annotations

Appropriate exposure

Appropriate compression

Absence of movement

Skin fold free

Absence of artefacts covering the image

Acceptability Criteria:

cranio-caudal

Medial border imaged

Some axillary tail shown

Pectoral muscle shadow may be shown

Nipple in profile

Correct annotations

Appropriate exposure

Appropriate compression

Absence of movement

Skin fold free

Absence of artefacts covering the image

Results

Of the 168 images, Experts 1, 2 and 3 considered 116, 155 and 101 respectively to be acceptable at the first read. At the second read, they classified 119, 151 and 118 as acceptable.

Within-expert agreement was reasonably high but ranged from under 80% to over 90% (Table 2).

Pairs of experts disagreed on whether an image was acceptable or not in at least 25% of cases (Tables 3 & 4).

Fleiss' kappa for agreement between all three experts was 0.24 at the first read and 0.339 at the second read.

Table 2: Intra-rater agreement

Expert 1	87.5 %	#=0.70
Expert 2	92.9%	#=0.56
Expert 3	78.0%	#=0.52

Table 3: Inter-rater agreement - first reading

Expert 1 versus Expert 2	73.2%	#=0.21
Expert 1 versus Expert 3	74.4%	#=0.45
Expert 2 versus Expert 3	64.3%	#=0.14

Table 4: Inter-rater agreement - second reading

Expert 1 versus Expert 2	75.0%	#=0.45
Expert 1 versus Expert 3	62.5%	#=0.56
Expert 2 versus Expert 3	67.2%	#=0.17

Conclusion

This work was embedded in a larger project to develop a training tool to improve decision-making on adequacy of clinical image quality, and thereby reduce excessive technical repeats in digital mammography. The training tool calls for a reference standard against which the trainees' decisions can be measured. Rather than taking a single expert opinion as the reference, aware of the potential for observer variability in this task and therefore questionable validity of a single opinion, we elected to develop a consensus categorisation of each of the images in the training and test pool. As part of the consensus-developing process among three expert observers, we assessed intra- and inter-observer variability. Our results appear to indicate that this was considerable, with no better than "Moderate" agreement between pairs of observers.

Our work adds to a very small existing body of literature attempting to quantify the reliability of clinical image quality assessment in mammography. A review in 2010 highlighted the wide range of scales which have been employed in mammography image quality research and the fact that many of them have not been subjected to rigorous reliability or validity testing[5]. Moreira et al in 2005[6] assessed the reliability of two scales which are in widespread clinical use - the 4-category PGMI system and the 3-category EAR (Excellent, Acceptable, Repeat) scale. The authors additionally dichotomised the scales into accept or reject, which produced inter-rater reliability percentages broadly similar to those in our study but with even lower kappa values.

It should be noted that the uneven distribution of the two categories in our sample (and the highly uneven distribution in that of Moreira et al) increases the risk of the crude percentage agreement level being affected by chance[7]. Furthermore, when there is uneven prevalence between categories and skewed agreement across categories, Cohen's kappa, designed to assess the extent to which agreement is greater than would be expected by chance, can produce paradoxical results[8]. Future work on observer variability in this context should perhaps use more evenly balanced samples, although this in turn risks creating a less lifelike exercise.

As well as observer factors, test reliability, or lack thereof -depends on subject variability and measurement error[7]. Subject variability is likely to be important when considering mammographic positioning - we would expect this to arise from anatomical variations. To improve clinical image quality scales for mammography, a better understanding of what is achievable in a given subject would be valuable. However, we suggest that measurement error resulting from the framing of the criteria in the scales is potentially highly important. For example, a criterion such as "appropriate compression" is very loosely defined and therefore likely to increase measurement error and observer variability due to varying interpretations of what is "appropriate".

Multiple psychological influences are likely to have influenced the observers' categorisation of mammography image quality in our study, and further work on observer variability should investigate psychological aspects. Our observers were all practicing or former trainers and competency assessors in mammography. One could therefore hypothesise that they might be prone to hypercritical quality assessments as part of habitually striving to instil high standards of practice in their students. In addition, the observers, although pseudonymised for the analysis, were acquainted with the principal investigator and, in most cases, each other. Another factor at play could therefore potentially have been fear of being judged not to hold high standards themselves. Further influences on the observers' scoring could include the difference between the study process and routine practice, and the observers' understandings of the aims of the exercise. Factors such as these could perhaps usefully be explored by in-depth qualitative interviews with observers.

Our work has expanded the body of quantitative evidence demonstrating the problem of observer variability in clinical image quality evaluation in mammography. This re-emphasises the need for greater standardisation, which could in turn be achieved by our computerised training tool. Automated image quality assessment methods are starting emerge[9] but until such time as they are fully developed and validated, human judgements will continue to be important.

A reference standard has been established for 168 images to be used in a computerised training tool to improve clinical image quality assessment in mammography. The reference standard consists of an overall classification for each image, derived from the majority decisions of six observations by three expert observers, with arbitration by a fourth expert when required. Establishing this

robust reference standard for the computerised training tool enables it to be taken forward to efficacy and effectiveness testing.

Personal information

Patsy Whelehan is a clinical research radiographer in breast imaging, with experience in mammography education and quality assessment.

Dr Mark Hartswood is a computer scientist whose research interests include human-computer interaction and health informatics.

Ann Mumby is the principal investigator on the full project to develop and implement an interactive training tool to help counteract subjectivity in clinical image quality assessment in mammography. She has been Mammography Training Co-ordinator for the Scottish Breast Screening Programme for 15 years.

The authors would like to thank

- the experts who reviewed the images for this project
- Petra Rauchhaus, for preliminary advice on the statistical analysis
- Professor Andy Evans, for reviewing the abstract.

We are also grateful to the Society and College of Radiographers, whose Industrial Partnership Scheme funded the project which encompasses this piece of work.

Images for this section:

Fig. 6

Fig. 5

Fig. 4

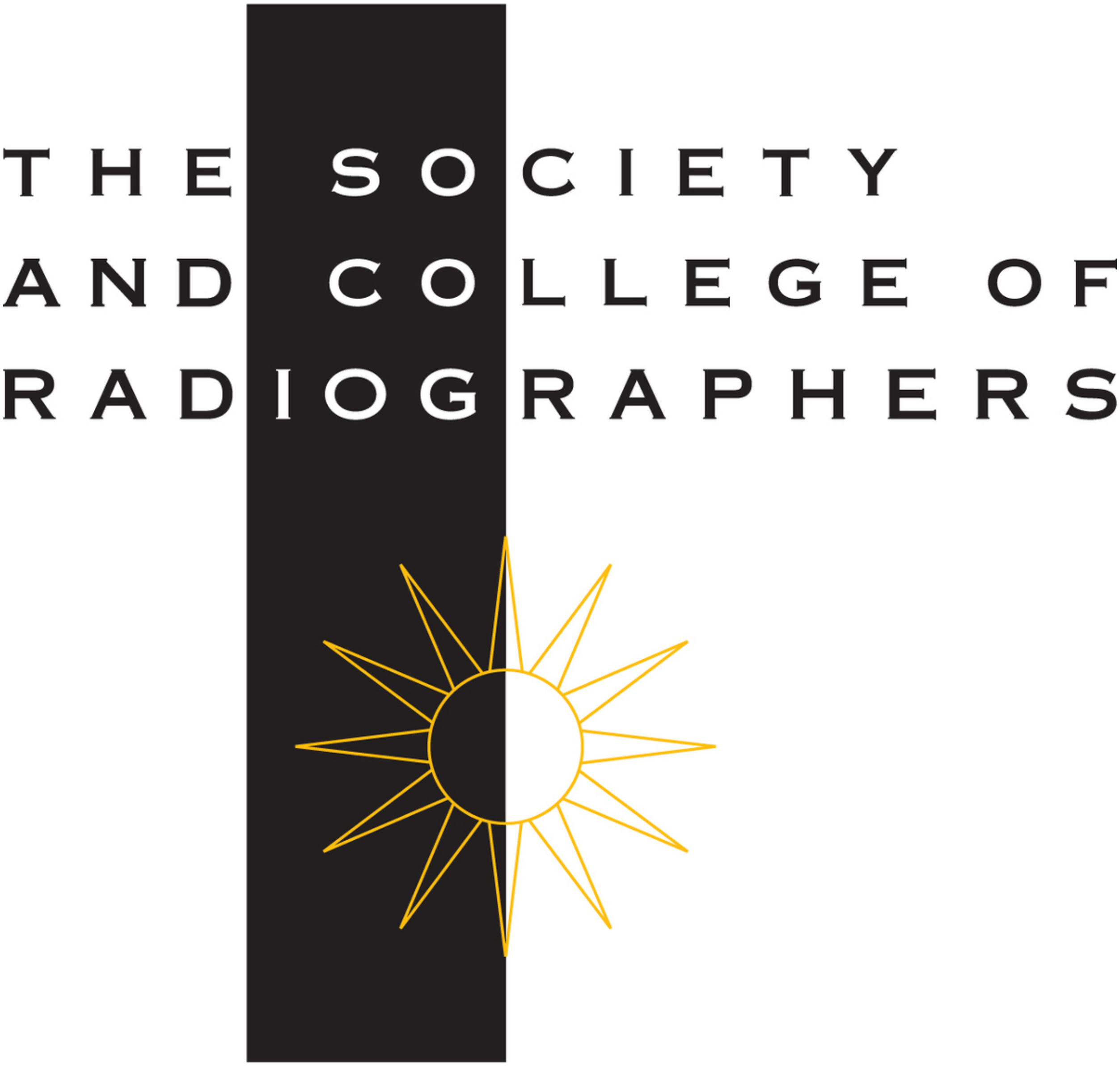


Fig. 7

References