ORIGINAL PAPER

Nagoya J. Med. Sci. **68**. 147 ~ 153, 2006

INTRA- AND INTEROBSERVER AGREEMENT AND PERFORMANCE SCORE OF BREAST PHANTOM IMAGE INTERPRETATION: INFLUENCE OF AMBIENT ROOM LIGHTING LEVELS

KAZUYUKI KOYAMA[1)], KAZUHIRO SHIMAMOTO[2)], MITSURU IKEDA[2)], HIDEYUKI MURAMOTO[3)], HIROKO SATAKE[4)], AKIKO SAWAKI[4)], KATSUHIKO KATO[4)], HIROMICHI FUKUSHIMA[4)] and TAKEO ISHIGAKI[4)]

[1)] *Faculty of Radiological Technology, Fujita Health University School of Health Sciences, 1-98 Dengakugakubo, Kutsukake-cho, Toyoake, Aichi 470-1192, Japan*
[2)] *Department of Radiological Technology, Nagoya University School of Health Sciences, 1-1-20 Daikominami, Higashi-ku, Nagoya 461-8673, Japan*
[3)] *Department of Radiology, Kasugai Municipal Hospital, 1-1-1 Takagi-cho, Kasugai, Aichi 486-8510, Japan*
[4)] *Department of Radiology, Nagoya University School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8560, Japan*

ABSTRACT

The influence of ambient room lighting conditions on soft-copy breast phantom image interpretation was evaluated by comparing cathode ray tube (CRT) monitors with liquid crystal displays (LCDs). Nine observers were asked to use a three-point scale to rate the visibility of various phantom objects (masses, specks, and fibers) displayed on a 21-inch CRT (2,560 × 2,048) and a 21-inch LCD (2,560 × 2,048) under three different levels of ambient lighting (20, 100 and 420 lux at the display center). Each phantom image was interpreted twice, and the reproducibility of judgment and inter-observer agreement was evaluated using kappa statistics. Except for the "mass" score, the LCD score showed a significantly higher value ($p<0.05$) compared with that of CRT. Nevertheless, no significant differences were found among the three lighting levels. Furthermore, intra- and inter-observer agreement in judgments showed no effects of room illumination. Although the breast phantom objects were better visualized on LCDs than on CRT monitors, room illumination did not affect the performance score of soft-copy reading.

Key Words: Mammography, Observer performance, Observer variation, Cathode ray tube (CRT) display, Liquid crystal display (LCD), Image interpretation

INTRODUCTION

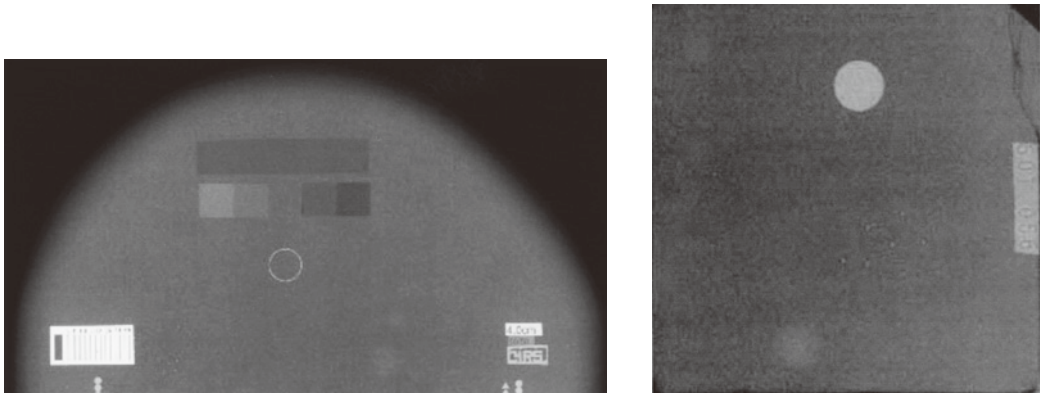
With recent advances in digital technology, full digital mammography systems have been introduced in clinical practice, and several investigators have reported on the technological requirements for mammographic displays.[1-3)] Preliminary results suggest that observer performance in soft-copy (monitor) reading is almost equal to that in hard-copy reading.[4)] Clinical mammograms require high-resolution and high-contrast images for the detection and characterization of

Corresponding author: Kazuhiro Shimamoto
Department of Radiological Technology, Nagoya University School of Health Science,
1-1-20 Daikominami, Higashi-ku, Nagoya 461-8673, Japan
Phone: +81-52-719-1562 Fax: +81-52-719-1509 / E-mail: simamoto@met.nagoya-u.ac.jp

microcalcifications, and image interpretation is performed in a dark room using high-luminance view boxes (at least 3,000 cd/m^2) with the help of a magnifying lens.[5] Similarly, in reading a cathode-ray tube (CRT) monitor, radiologists prefer a dark room in order to avoid reductions in image contrast due to the reflection of ambient lighting. Although liquid crystal displays (LCDs) are gradually replacing CRTs for soft-copy medical image interpretation, the characteristics of LCDs such as angular response, are quite different from those of CRTs, and the optimal reading environment for a high-luminance LCD has not yet been established. It is essential to clarify the optimal reading environment in which soft-copy reading becomes an acceptable replacement for screen-film mammogram interpretation. In the present study, we compared CRTs with LCDs in evaluating the influence of ambient room illumination on the scoring system used in breast phantom image interpretation.

MATERIALS AND METHODS

A CIRS model 12A (Computerized Imaging Reference System, Inc., Norfolk, VA, USA) and RMI model 156 (Gammex RMI, Middletown, WI, USA) were used in the present study (Fig. 1). The CIRS model 12A includes seven embedded simulated masses (75% glandular, 25% adipose; range, 0.90 to 4.76 mm in thickness), twelve groups of calcification specks (calcium carbonate; range, 0.13 to 0.40 mm in diameter), and five fibers (nylon fiber; range, 0.30 to 1.25 mm) as the test objects; the RMI model 156 has six embedded simulated masses (phenolic rod; range, 0.25 to 2 mm in thickness), five groups of calcification specks (aluminum oxide; range, 0.16 to 0.54 mm in diameter), and five fibers (nylon fiber; range, 0.40 to 1.56 mm). At first, screen-film phantom images were obtained using a Kodak MIN-R EV screen-film and a commercially available mammography unit, MAMMOMAT3000 (Siemens-Asahi Meditec, Co., Tokyo, Japan) with an exposure of 27kVp using the automatic exposure control (AEC). Three film images were obtained for each phantom, the best of which was selected for the study. In order to generate Digital Imaging and Communications in Medicine (DICOM) images, the screen-film images were then digitized using an LD5500 film digitizer (Konica Minolta, Inc., Tokyo, Japan) with a sampling pitch of 50 micrometers and a gray-scale contrast resolution of 12 bits. For soft-copy image interpretation, a 21-inch CRT monitor with a matrix resolution of 2,048 × 2,560



(a) Model 12A (b) Model 156

Fig. 1 Breast phantom image.

(MDG521; Barco, Kortrijk, Belgium) and a 21.3-inch LCD with a matrix resolution of 2,048 × 2,560 (RadiForce® G51; Nanao, Co., Ishikawa, Japan) were used. The gray-scale resolution was 8 bits for both CRT and LCD, and 12-bit image data were displayed by reducing each 12-bit pixel value to 8-bit. The maximum luminance of each monitor was calibrated to 450cd/m^2 at the monitor center. The minimum luminance was set to 0 cd/m^2 for CRT, and 1cd/m^2 for LCD. To minimize the glare due to reflected light, a CRT screen uses an antireflection (AR) coat whereas an LCD screen is treated with antiglare (AG) coatings. The DICOM Image viewing software for both displays was RS252DV (Konica Minolta, Inc., Tokyo, Japan). In order to display the same gray-scale contrast as the original films, a linear output of digitized images was employed, and none of the image-processing functions (such as the change of contrast and brightness setting of the displays) were allowed. The size of the displayed phantom images was fitted to the display screen size. Room illuminance was measured using a digital illuminance meter (IM-3; Topcon, Tokyo, Japan), and was adjusted to 20, 100 and 420 lux at the center of each display (20, 120, 480 lux at the console desk, respectively). These ambient lighting conditions were identical to those used in our previous studies.[6,7] Three types of reading sessions with different ambient lighting levels were conducted for each observer, and the order of the reading sessions was randomized.

Nine observers (seven radiologists and two radiological technologists, each with over ten years of clinical experience) were asked to determine the visibility of each phantom test object. No time limit was imposed for reaching a judgment. The observers used a three-point (1, 0.5 and 0) scale based on the American College of Radiology (ACR) standards.[8] For the "mass" score, each mass was assigned a rating of 1 if it was clearly seen in the correct location; 0.5 if it was visible but lacked a generally circular appearance; and 0 if it was not seen. For the "speck" score, the number of visible specks for each speck group was recorded. Each speck group as a whole was assigned a rating of 1 if four or more specks were visible; 0.5 if two or three were visible; and 0 if none or only one was visible. For the "fiber" score, each fiber was assigned a rating of 1 if the full length of the fiber was visible; 0.5 if at least half but not all of the fiber was visible; and 0 if only half or less than half was visible. Finally, the performance score was defined as the total score for each test object. To carry out the statistical analysis of reproducibility of judgments (intra- observer agreement), Cohen's kappa value[9,10] was employed. Such a value ranging from 0 to 0.20 was regarded as poor agreement; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, good; and 0.81–1.0, excellent. Similarly, inter-observer agreement among the nine observers was evaluated using the kappa statistics described by Fleiss.[11]

RESULTS

The average performance scores for each monitor are summarized in Table 1. In both phantom images, analysis of variance (ANOVA) showed that, except for the "mass" score, LCD scores with or without the protective cover were significantly higher than CRT scores ($p<0.05$). Similarly, the actual number of specks visible on LCDs was significantly higher than that on CRT monitors ($p<0.01$) (Table 2). However, we found no dependence of performance score on the levels of room illumination.

Reproducibility of judgment is summarized in Table 3. Under all viewing conditions, the kappa value was greater than 0.40 (moderate agreement or more), and no significant differences in kappa values were found between the CRT and LCD interpretations. As shown in Table 4, there were no significant differences in kappa values among the three kinds of room lighting levels, although differences in kappa values were noted among the different types of objects.

Table 1 Average score comparison under three kinds of ambient lighting conditions.

Object	Type of monitor[a]	Model 12A phantom Full score	20 lux	120 lux	480 lux	
mass	CRT	7.00	6.10	5.95	6.15	
	LCD (+)	7.00	6.10	5.95	6.10	
	LCD (–)	7.00	6.20	6.20	6.33	
speck	CRT	12.0	11.48	11.40	11.30	[b]
	LCD (+)	12.0	11.55	11.53	11.58	[b]
	LCD (–)	12.0	11.65	11.55	11.65	
fiber	CRT	5.00	4.13	3.95	3.83	[b]
	LCD (+)	5.00	4.10	4.18	4.13	[b]
	LCD (–)	5.00	4.25	4.25	4.28	
Total score	CRT	24.0	21.70	21.30	21.28	
	LCD (+)	24.0	21.75	21.65	21.80	[b]
	LCD (–)	24.0	22.10	22.00	22.18	

Object	Type of monitor [a]	Model 156 phantom Full score	20 lux	120 lux	480 lux	
mass	CRT	5.00	4.10	4.01	4.10	
	LCD (+)	5.00	4.10	4.01	4.00	
	LCD (–)	5.00	4.15	4.00	4.05	
speck	CRT	5.00	3.53	3.35	3.35	[b]
	LCD (+)	5.00	3.70	3.75	3.63	[b]
	LCD (–)	5.00	3.78	3.73	3.75	
fiber	CRT	6.00	4.93	4.75	4.78	[b]
	LCD (+)	6.00	5.15	5.23	5.15	[b]
	LCD (–)	6.00	5.30	5.25	5.28	
Total score	CRT	16.00	12.55	12.15	12.23	[b]
	LCD (+)	16.00	12.95	13.08	12.78	[b]
	LCD (–)	16.00	13.23	12.98	13.08	

[a] For LCD, (+)=with protective cover; (–)=without protective cover.
[b] ANOVA showed a significant difference between CRT and LCD ($p<0.05$).

Table 2 Average total number of visible specks under three kinds of ambient lighting conditions.

Phantom type	Type of monitor[a]	Total number of specks	Ambient lighting conditions 20 lux	120 lux	480 lux	
Model 12A	CRT	144	135.8	133.6	133.2	[b]
	LCD (+)	144	137.8	136.5	137.6	[b]
	LCD (–)	144	138.7	138.3	138.9	
Model 156	CRT	60	40.1	40.2	40.1	[b]
	LCD (+)	60	42.4	43.6	42.6	[b]
	LCD (–)	60	43.7	42.1	43.5	

[a] For LCD, (+)=with protective cover; (–)=without protective cover.
[b] ANOVA showed a significant difference between CRT and LCD ($p<0.05$).

Table 3 Reproducibility of judgment under three kinds of ambient lighting conditions.

Object	Type of monitor[a]	Model 12A 20 lux	120 lux	480 lux	Model 156 20 lux	120 lux	480 lux
mass	CRT	0.577	0.618	0.636	0.800	0.900	1.000
	LCD (+)	0.665	0.687	0.473	0.850	0.955	0.855
	LCD (–)	0.536	0.736	0.770	0.900	0.855	0.900
speck	CRT	0.873	0.658	0.658	0.704	0.832	0.812
	LCD (+)	0.800	0.812	0.610	0.877	0.839	0.775
	LCD (–)	0.848	0.627	0.748	0.819	0.875	0.770
fiber	CRT	0.264	0.387	0.475	0.817	0.494	0.512
	LCD (+)	0.412	0.209	0.277	0.757	0.639	0.609
	LCD (–)	0.373	0.409	0.332	0.814	0.709	0.638

[a] For LCD, (+)=with protective cover; (–)=without protective cover.

Table 4 Inter-observer agreement under three kinds of ambient lighting conditions.

Object	Type of monitor[a]	Model 12A 20 lux	120 lux	480 lux	Model 156 20 lux	120 lux	480 lux
mass	CRT	0.741	0.737	0.780	1.000	1.000	0.781
	LCD (+)	0.421	0.567	0.471	0.795	0.795	0.835
	LCD (–)	0.670	0.439	0.388	0.781	0.835	1.000
speck	CRT	0.446	0.476	0.458	0.650	0.696	0.718
	LCD (+)	0.430	0.460	0.432	0.632	0.633	0.697
	LCD (–)	0.404	0.485	0.396	0.632	0.637	0.689
fiber	CRT	0.539	0.568	0.635	0.491	0.529	0.605
	LCD (+)	0.604	0.655	0.650	0.421	0.478	0.360
	LCD (–)	0.584	0.587	0.470	0.377	0.457	0.358

[a] For LCD, (+)=with protective cover; (–)=without protective cover.

DISCUSSION

In clinical practice, daily evaluations of breast phantom images are routinely performed for the quality control of mammograms, since the accurate assessment of phantom objects is essential in maintaining the image quality of clinical mammograms. To avoid inter-observer variance, all phantom images should be judged under identical viewing conditions, including the same ambient lighting levels and by the same person who should have solid experience in mammogram interpretation.[5,8] Although computer-aided detection of phantom objects has been developed,[12] human observation should still be required. However, the scoring of phantom images requires subjective judgment, and different observers can see different numbers of test objects in the same image. The present results suggest that there is no significant effect of room lighting levels and/or type of monitor on intra- and inter-observer agreements in judgments. Compared with model 156, the kappa values of model 12A tended to be lower in "mass" and "speck" interpretations. In these test objects, the materials of the objects were different between the two phantom models, and the number of objects detected in model 12A was higher than that in model 156. Additionally, model 12A contained smaller specks (0.13 mm in diameter) than model 156, a fact which could amplify intra- and inter-observer variances.

Room lighting levels are a key issue in mammography interpretation. In the present study, to minimize the impact of effects other than lighting levels, the observers were not permitted to use any of the image-processing functions. During monitor reading, the glare caused by light reflected from the display screen can degrade observer performance. However, our results suggest that the influence of room lighting levels is not obvious in CRT and LCD interpretation when using a current high-resolution and high-contrast monitor. Scharoizer *et al.*[13] report that the detection performance of catheters on bedside chest radiographs with both CRT monitors and LCDs was equally reduced by bright ambient light (> 100 lux, in front of the monitor). However, in their study, the maximum luminance of the monitors had been adjusted to 300 cd/m^2, whereas the present monitor luminance of 450 cd/m^2 was 50% higher. Additionally, our previous work[6] reported that only the combination of high room illuminance (480 lux) and low CRT luminance (50 cd/m^2) significantly degraded the detectability of pulmonary nodules ($p<0.05$) under nine combined conditions of CRT luminance (50, 200 and 500 cd/m^2) and room illuminance (20, 120 and 480 lux). Therefore,we believe that the inconsistency of the results obtained by Scharoizer *et al.*[13] is not critical.

Krupinski *et al.*[14] report that observer viewing with LCD displays is superior to that with CRT monitors, at least in on-axis viewing. Consistent with their study, our present results show that the viewing score with LCD displays is superior to that obtained with CRT monitors. Although the differences in actual scores were very small (less than 1.0 point), they were significant from the viewpoint of the quality control of mammograms even at only a one-rank difference on a three-point rating scale in only one object. The superiority of LCD over CRT can be explained by differences in the physical characteristics of the displays such as their modulation transfer function (MTF) and veiling glare.[14] The type of antireflective techniques for the reduction of glare differs between CRT and LCD, with the reflected glare of a CRT screen being greater than that of an LCD.

In conclusion, the effect of room illuminance in the evaluation of breast phantom images is not significant when using a current high-luminance monitor. Furthermore, LCDs can provide better visualization of objects than CRT monitors.

ACKNOWLEDGEMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS) (Project No. 16591200).

REFERENCES

1) Hemminger, B.M.: Soft copy display requirements for digital mammography. *J. Digit. Imaging*, 16, 292–305 (2003).
2) Samei, E. and Wright, S.L.: Luminance and contrast performance of liquid crystal displays for mammographic applications. *Technol. Cancer Res. Treat.*, 3, 429–436 (2004).
3) Samei, E.: AAPM/RSNA physics tutorial for residents: technological and psychophysical considerations for digital mammographic displays. *Radiographics*, 25, 491–501 (2005).
4) Obenauer, S., Herman, K.P., Marten, K., Luftner-Nagel, S., von Heyden, D., Skaane, P. and Grabbe, E.: Soft copy versus hard copy reading in digital mammography. *J. Digit. Imaging*, 16, 341–344 (2003).
5) Castaneda, G.: Phantom images. *In Breast Imaging Companion*, 2nd edition. pp. 37–40 (2001), Lippincott Williams & Wilkins, Philadelphia.
6) Ishihara, S., Shimamoto, K., Ikeda, M., Kato, K., Mori, Y., Ishiguchi, T. and Ishigaki, T.: CRT diagnosis of pulmonary disease: influence of monitor brightness and room illuminance on observer performance. *Comput. Med. Imaging Graph.*, 26, 181–185 (2002).
7) Muramoto, H., Shimamoto, K., Ikeda, M., Koyama, K., Fukushima, H. and Ishigaki, T.: Influence of monitor luminance and room illumination on soft-copy reading evaluation with electronically generated contrast-detail phantom: comparison of cathode-ray tube monitor with liquid crystal display. *Nagoya J. Med. Sci.* (in press).
8) American College of Radiology: Mammography quality control manual. Reston, American College of Radiology (1999).
9) Cohen, J.A.: A coefficient of agreement for nominal scales. *Educat. Physiol. Meas*, 20, 37–46 (1960).
10) Cohen, J.A.: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70, 213-220 (1968).
11) Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76, 378–382 (1971).
12) Jiang, Y., Nishikawa, R.M., Schmidt, R.A., Toledano, A.Y. and Doi, K.: Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications. *Radiology*, 220, 787–794 (2001).
13) Scharoizer, M., Prokop, M., Weber, M., Fuchsjager, M., Oschatz, E. and Schaefer-Prokop, C.: Detectability of catheters on bedside chest radiographs: comparison between liquid crystal display and high-resolution cathode-ray tube monitors. *Radiology*, 234, 611–616 (2005).
14) Krupinski, E.A., Johnson, J., Roehrig, H., Nafziger, J., Fan, J. and Lubin, J.: Use of a human visual system model to predict observer performance with CRT vs. LCD display of images. *J. Digit. Imaging*, 17, 258–63 (2004).