Journal of Digital Imaging

Comparison of Human Observer Performance of Contrast-Detail Detection Across Multiple Liquid Crystal Displays

Alice N. Averbukh, MS,[1] David S. Channin, MD,[1] and Prasobsook Homhual, PhD[2]

Appropriate selection of a display subsystem requires balancing the optimization of its physical parameters with clinical setting and cost. Recent advances in Liquid Crystal Display (LCD) technology warrant a rigorous evaluation of both the specialized and the mass market displays for clinical radiology. This article outlines step two in the evaluation of a novel 9.2 million pixel IBM AMLCD panel. Prior to these experiments, the panel was calibrated according to the DICOM Part 14 standard, using both a gray-scale and a pseudo-gray scale lookup table. The specific aim of this study is to compare human, contrast-detail perception on different computer display subsystems. The subsystems that we looked at included 3- and 5-million pixel "medical-grade" monochrome LCDs and a 9.2-million pixel color LCD. We found that the observer response was similar for these three display configurations.

KEY WORDS: LCD, phantom, contrast-detail, observer, perception

PHANTOMS ARE USED widely in radiology to evaluate human contrast-detail perception with different film-based modalities. This article describes limited assessment of human observer contrast-detail response with both desktop color and clinical monochrome LCDs. A digital contrast-detail phantom, designed to test human contrast-detail perception on electronic display systems, was employed for this task. The display of medical images in a diagnostic setting is a complex phenomenon. Hardware and software requirements as well as other human factors, such as viewing environment and ergonomics, all play a role in the choice of appropriate equipment. One key component of the viewing workstation is the display subsystem.

The display subsystem consists of a video card (also known as a frame buffer) and a display device. The latter is typically a cathode ray tube (CRT), or a flat panel display, usually a liquid crystal display (LCD), though this term can be abused in lay discussion to include a number of different flat panel display technologies.

Appropriate selection of a display subsystem requires balancing the optimization of the physical parameters of the display subsystem with clinical setting and cost. In a high-volume setting, where diagnostic accuracy is of prime importance, more weight might be given to maximizing physical performance at the expense of cost. In a lower volume image review setting, perhaps in a poorer ergonomic environment (e.g., a brightly lit nursing unit), a lower performance, and therefore cheaper, display subsystem might be appropriate. In most cases diagnostic workstations employ specialized high-resolution gray-scale displays with luminance ranges up six times greater than the luminance ranges of their commercial counterparts. Other improvements such as built-in photometers, specialized high-end graphics controllers, as well as specialized software optimized for these displays widens the gap with respect to the commercially available display technology [1-3]

However, recent advances in commercial LCD technology have brought about significant

[1]*From the Northwestern University Medical School, 448 E. Ontario Suite 300, IL, 60611, USA.*

[2]*From the University of the Thai Chamber of Commerce, 126/1 Vibhavadi Rangsit Road, 10320, Bangkok, Thaliland.*

Correspondence to: Alice N. Averbukh, MStel: 312-92601573; fax: 312-926-4220; e-mail: aaverbukh@radiology.northwestern.edu

Online publication 12 January 2005

doi: 10.1007/s10278-004-1035-1

improvements in their luminance, spatial resolution, spatial noise, and angular dependence, while lowering their cost.[4] Color LCDs have become an attractive desktop option. Validating such displays for diagnostic imaging would allow taking advantage of not only their color display properties but their lower mass market costs. While physical discrepancies are often measurable between mass market display devices and high-end "medical-grade" displays, the impact of these physical differences on clinical performance (accuracy of interpretation) must be assessed. A number of studies have been performed to verify mass market displays, though none, to our knowledge, yielded any statistically significant results.[5] In other words, it remains unclear whether the use of a mass market display subsystem with measurable deficiencies, compared to a high-end display, affects clinical outcome. This is analogous to the situation from the past, wherein film/screen radiography was compared to digital projection radiography. The physics of film/screen was in many aspects superior to digital radiography, but several studies showed that these physical differences did not affect clinical outcome in a number of difficult imaging scenarios [6-10]

Clinical evaluations, typically involving receiver operating characteristic (ROC) methodologies are long, difficult, and expensive, but in the absence of strong scientific data, purchasers of equipment are making decisions based on anecdotal information and marketing material.

MATERIALS AND METHODS

Three different LCD display configurations were compared. DOME C3 with a DOME Dx/PCI-2 graphics card (PLANAR Systems, Beaverton, OR), DOME C5i with a DOME Dx/PCI-2 graphics card (PLANAR Systems), and the IBM T221 (IBM Corp., Armonk, NY) driven by ATI Fire GL4 graphics card (ATI Technologies Inc., Ontario, Canada) with DICOM 3.14[9] pseudo-gray-scale (RGB) calibration. DICOM Part 14[11] calibration was performed on all three monitors. We have previously demonstrated that successful DICOM Part 14 calibration can be performed on some of these mass market displays.[12,13] Some common characteristics of these displays as well as the calibration results, wherever applicable, are juxtaposed in Table 1. In the course of the experiment, the ambient lights were set to be $\frac{1}{4}$ of the minimum luminance for a particular monitor for each session.

An IRB approval was obtained to enroll normal human volunteers and have their visual acuity tested by a board certified ophthalmologist. All observers had corrected acuity of 20/20. A total of 350 different test patterns were displayed to five observers in a randomized order on each display. The observers were asked to identify and rate the test patterns according to how obvious those appeared to them, on the scale from 0 to 5 (0 = no pattern observed; 5 = pattern is very obvious). To minimize observer fatigue each of the 1-h sessions has been broken down into 15-min periods with a 5-min break between periods.

The retinal viewing angle, subtended in the horizontal direction, was fixed at 78 degrees for our target of a fixed width of 50 pixels. This allowed us to fix the viewing distance for all observers so that the distance to T221 had to stay at 18 in., as recommended by the manufacturer (Steven Wright, IBM, personal communication). Thus, for our target of a fixed width of 50 pixels, the width in centimeters varied based on the resolution of the display (see Table 2)

Although IBM had to resign to the fairly close viewing distance of 18 in. due to the small pixel size of this display model, this distance is not considered ergonomically ideal, and larger pixel size would be preferable. The resting point of vergence (RPV) is the distance at which the eyes are set to converge when there is no object to converge on. This distance "averages at approximately about 45 inches when looking straight ahead and comes in to about 35 inches with a 30-degree downward gaze angle".[14] It is recommended that the viewing distance to a computer monitor be set no closer than the RPV distance to reduce eye strain. To maintain the constant viewing angle, the distance to C5i and C3 was fixed at 24 and 30 in., respectively, for all observers.All patterns of type 1 (Fig. 1) contain a 50% gray square background and a small circle in the center which is several shades of gray away from the background. Each observer would see 100 patterns of this type (60 positive, 40 negative) on each display. "Positive finding" here indicates a second circle, of the

Table 1. Display Characteristics, Luminance Measurements, and DICOM part 14 Calibration Results

	Resolution	Dot Pitch mm	DPI	L_{max}	L_{min}	Dyn Rng	Mean JND step	STD Dev %	Ave. delta JND	Max delta JND	Distinct gray levels
DOME C3	1536 × 2	0.207	123	647	1.25	518	2.6	13	0.01	3.51	256
DOME C5i	2048 × 2	0.165	154	635	1.05	605	2.62	43	0.05	5.76	256
IBM T221 P calibration	3840 × 2	0.125	203	245	0.99	248	2.07	16	0.02	2.91	256

Table 2. Distance, Viewing Angle Constraints Imposed for all Observers, and the Just Noticeable Difference in Contrast × Spot Size Value

	Target width in pixels	Target width, cm	Distance, cm	Angle	JND in CS units
DOME C3 GS calibra	50	1.035	75.7	78	+40/-40
DOME C5i GS calibra	50	0.825	60.3	78	+60/-40
IBM T221, PGS calib	50	0.625	45.7	78	+50/-50

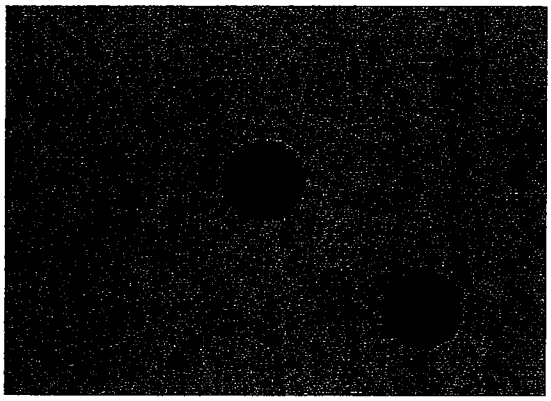


Fig 1. First, contrast-detail, pattern (enhanced for print publication); statistically significant differences in sensitivity observed between: C3 and C5i, C3 and T221.

same contrast as the one in the middle, located in one of the corners. A negative finding implies that only the middle circle is present. The observers had to identify the corner in which the second circle would appear and specify their confidence that the second circle was present. Both, the size of the circles, as well as the contrast with respect to the background, varied as summarized in Figure 2.

All patterns of type 2 (Fig. 3) contain a 50% gray square background and a set of 2 small vertical and 2 small horizontal lines of varying thickness and contrast, representing a cross. Each observer would see 50 (20 positive and 30 negative) of these patterns. "Positive finding" here indicates which one of the 4 small lines differs in contrast from the other three. A "negative finding" implies that all 4 lines are the same. The width and contrast of the lines varied as summarized in Fig. 4.

We also introduced a set of patterns to test gray-scale grouping, based on the Gestalt approach to human perception and problem solving.[15-17] Thus the third pattern (Figure 5) was created to resemble colorblindness test patterns in which the observer must identify a pattern made up of colored circles. In our case the pattern contained varying size/contrast monochrome circles representing an "E," hidden among other monochrome "background" circles set within a 78% gray square. The observers had to detect the letter "E" and determine the direction it was facing for 200 (100 positive / 100 negative) patterns of this type. This tested human perception of gray scale grouping. Correctly identifying the letter by choosing the direction it is facing constitutes a "positive finding" here, whereas no letter indicates a "negative finding." Thus the pattern was made up of monochrome circles whose radius was randomly chosen and fell in the range of 1 to 15 pixels. The contrast, as the difference in DDLs between background circles and the foreground ones forming the letter "E" went from 0 to 30, where the background DDL was in the range between 28 and 228.

RESULTS

There were three comparisons performed for each pattern. If no Bonferroni correction [19] was

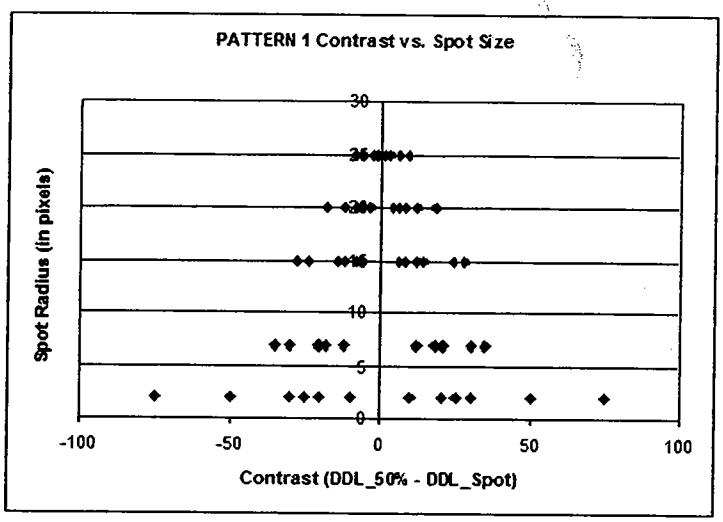


Fig 2. Pattern 1: Contrast as delta DDL vs. spot size radius in pixels.

applied, the probability of finding one or more significant differences by chance alone is 0.1426 (14.26%). Thus we have to lower the alpha for each test to 0.016952428 to bring the alpha level overall back to 0.05. Because having multiple comparisons reduces the statistical power of this experiment, the results below will indicate *P* values lower then 0.05 to indicate possible trends. For each comparison, a *P* value above 0.016952428 would not constitute statistical significance.

Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value

The sensitivity, specificity, positive predictive value, and negative predictive value data for the detection of the three patterns on each of the displays are shown in Figs. 6, 7, and 8, respectively. The specificity data was nearly perfect for all of the patterns on all of the displays. In the absence of spatial noise, observers did not perceive any ghost findings when patterns were displayed to them, as indicated by the perfect specificity scores.

For each pattern, we have performed pair-wise sensitivity comparisons using the McNemar's test of correlated proportions to assess statistical significance of any differences in sensitivity between the various display configurations.

For all patterns, significant differences in sensitivity were observed between C3 and each

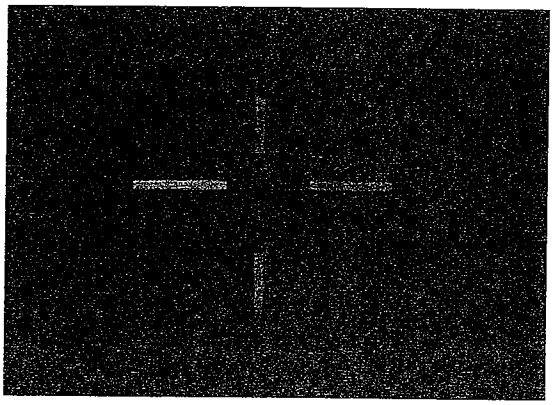


Fig 3. Second, contrast-detail, pattern (enhanced for print publication); no statistically significant differences were observed.

the C5i and T221 ($P < 0.001$). No statistically significant differences were found between C5i and IBM T221 configurations. C3 had the best performance in terms on sensitivity for the patterns of type 1. C5i performed best for the patterns of type 2, and T221 configuration had the highest sensitivity among the three for the patterns of type 3, although no statistically significant differences were found for patterns of type 2 and 3.

Sensitivity values were highest for the patterns of type 2, suggesting that these patterns were easier to identify. The grouping task (pattern type 3) was by far the most difficult

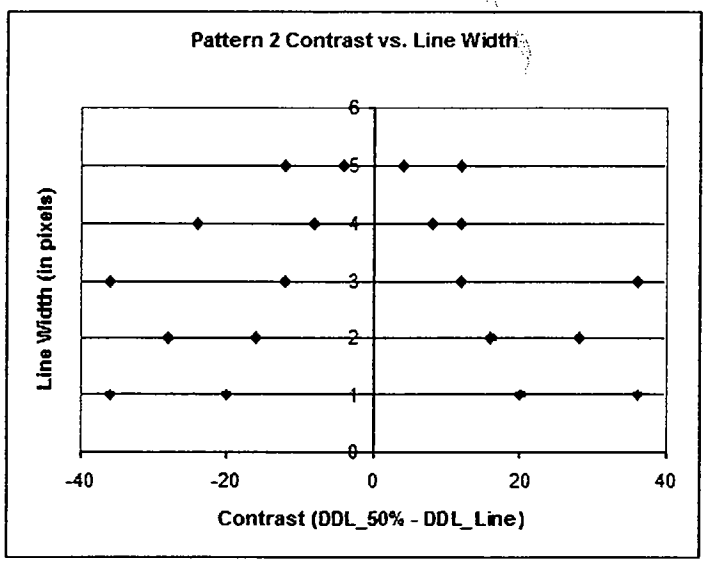


Fig 4. Pattern 2: Contrast as delta DDL vs. line width in pixels.

among the three, as indicated by the overall lower sensitivity values. Observer sensitivity here was slightly lower for the C3 configuration than for the other two display configurations.

Observer Confidence and Inter-Observer Variability

The observer confidence values for each of the patterns are shown in Figs. 9, 10, and 11, respectively. A single factor ANOVA test of statistical significance was applied for pair-wise comparison. No statistically significant differences were observed between the display configurations, although the overall observer confidence was slightly better for the C5i configuration for all pattern types.

For all pattern types we found a strong correlation between observer responses with little inter-observer variability:

For patterns of type 1 the Average Measure Intraclass Correlation = 0.9556; 95.00% C.I.: Lower = 0.947; Upper = 0.9631.
For patterns of type 2 the Average Measure Intraclass Correlation = 0.831; 95.00% C.I.: Lower = 0.778; Upper = 0.875.
For patterns of type 3 the Average Measure Intraclass Correlation = 0.965; 95.00% C.I.: Lower = 0.960; Upper = 0.969.

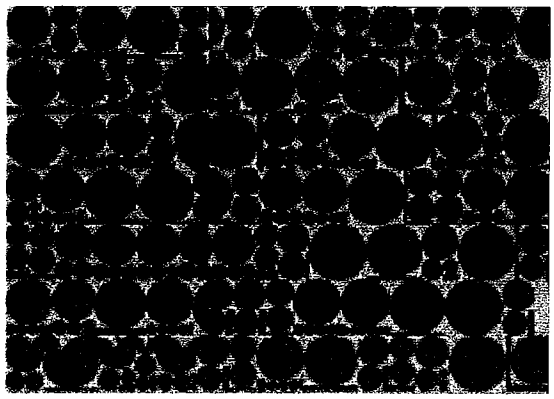
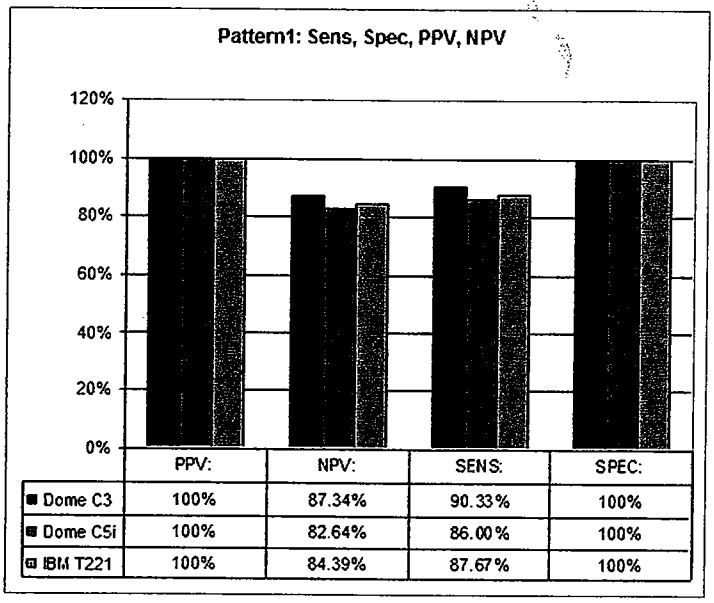


Fig 5. Third, gray scale grouping, pattern (enhanced for print publication); no statistically significant differences were observed.

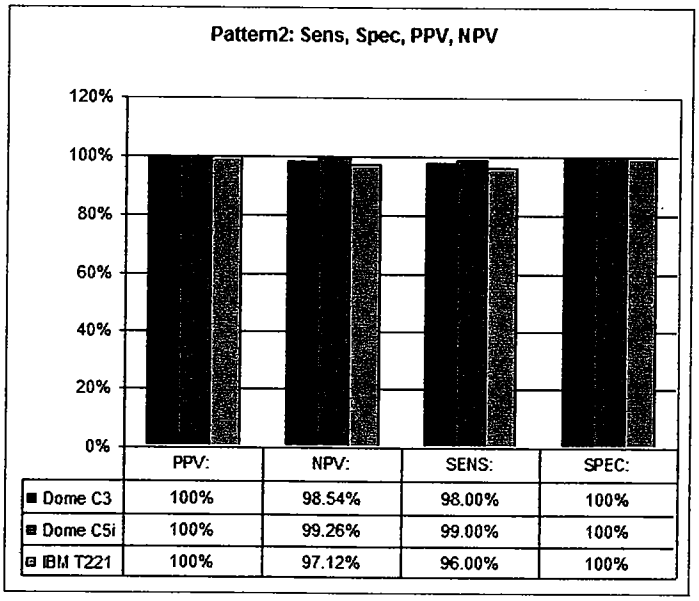
Contrast / Noise

Patterns of type 1 provide a good way to compare contrast/noise characteristics of each display configuration by calculating the just noticeable difference in the "contrast"*"spot size" product (JNCS), where contrast of the spot is measured as the difference in DDLs with respect to a 50% gray background, and the "spot size" is the radius of the spot, in pixels. The "contrast," and therefore the CS product, is a positive value when the foreground shade of gray is brighter than the 50%



	PPV:	NPV:	SENS:	SPEC:
Dome C3	100%	87.34%	90.33%	100%
Dome C5i	100%	82.64%	86.00%	100%
IBM T221	100%	84.39%	87.67%	100%

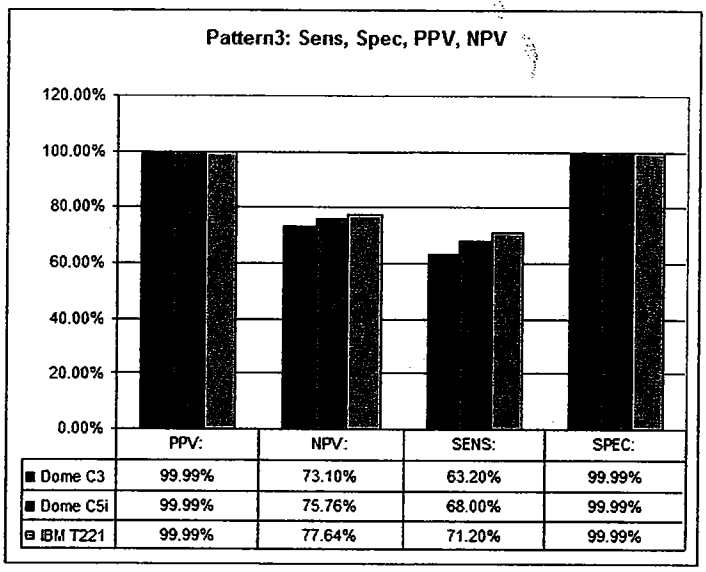
Fig 6. PPV, NPV, sensitivity, and specificity for Pattern 1.



	PPV:	NPV:	SENS:	SPEC:
Dome C3	100%	98.54%	98.00%	100%
Dome C5i	100%	99.26%	99.00%	100%
IBM T221	100%	97.12%	96.00%	100%

Fig 7. PPV, NPV, sensitivity, and specificity for Pattern 2.

background. The CS product is negative where the gray shade of the "finding" is darker than the 50% background. The JNCS for 75% correct with respect to a 50% gray background was used as a quantitative metric as described in Table 2. Fig. 12 juxtaposes observed conspicuity as a function of CS for the DOME C3, DOME C5i, and the RGB calibrated IBM T221 displays. The results suggest that there are insignificant differences in luminance noise between the three display combinations. The curve for IBM T221 is smoother, suggesting a more consistent response for all confidence levels on this display.



	PPV:	NPV:	SENS:	SPEC:
Dome C3	99.99%	73.10%	63.20%	99.99%
Dome C5i	99.99%	75.76%	68.00%	99.99%
IBM T221	99.99%	77.64%	71.20%	99.99%

Fig 8. PPV, NPV, sensitivity, and specificity for Pattern 3.

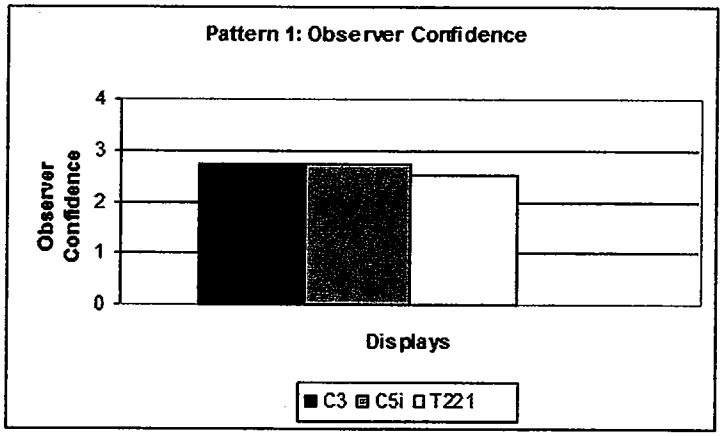


Fig 9. Observer confidence for Pattern 1; no statistically significant differences in observer confidence between three displays.

Time Spent per Pattern

The time it took each observer to identify and rate a pattern was also tracked. We explored the correlation between the time spent per pattern and both the sensitivity and observer confidence. Polynomial regression analysis was carried out to assess any correlation and its statistical significance.

We found a strong positive correlation between time spent per pattern and both the observer sensitivity and observer confidence for patterns 1 and 2 (Figs. 13a, 13b and 14a, 14b, respectively) We found no significant correlation between time and observer confidence for patterns of type 3 (Fig. 15b), and a negative correlation between time and sensitivity (Fig. 15a) for this pattern type.

DISCUSSION

There were significant but small differences in observer sensitivity, and no significant differences in observer confidence and contrast-noise characteristics for the three display configurations. This suggests to us that the three

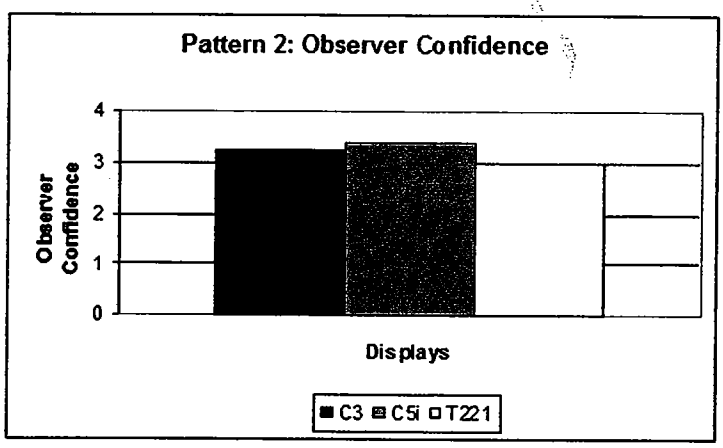


Fig 10. Observer confidence for Pattern 2; no statistically significant differences in observer confidence between three displays.

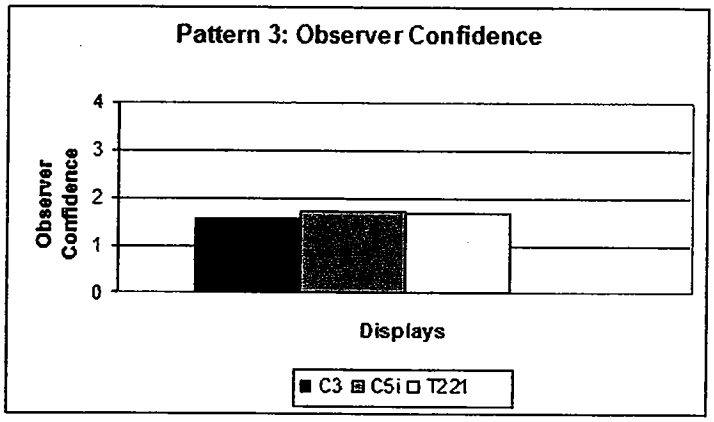


Fig 11. Observer confidence for Pattern 3; no statistically significant differences in observer confidence between three displays.

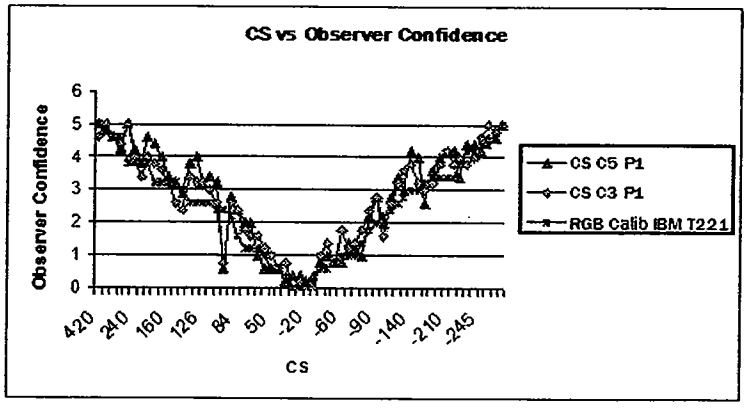


Fig 12. "Contrast" × "Spot Size" product (CS) vs. Observer Confidence; The "Contrast" is measured as delta DDLs with respect to a 50% gray background, with positive values indicating higher relative DDLs, and negative values representing lower (darker) DDLs. The "Spot Size" indicates radius of the spot in pixels.

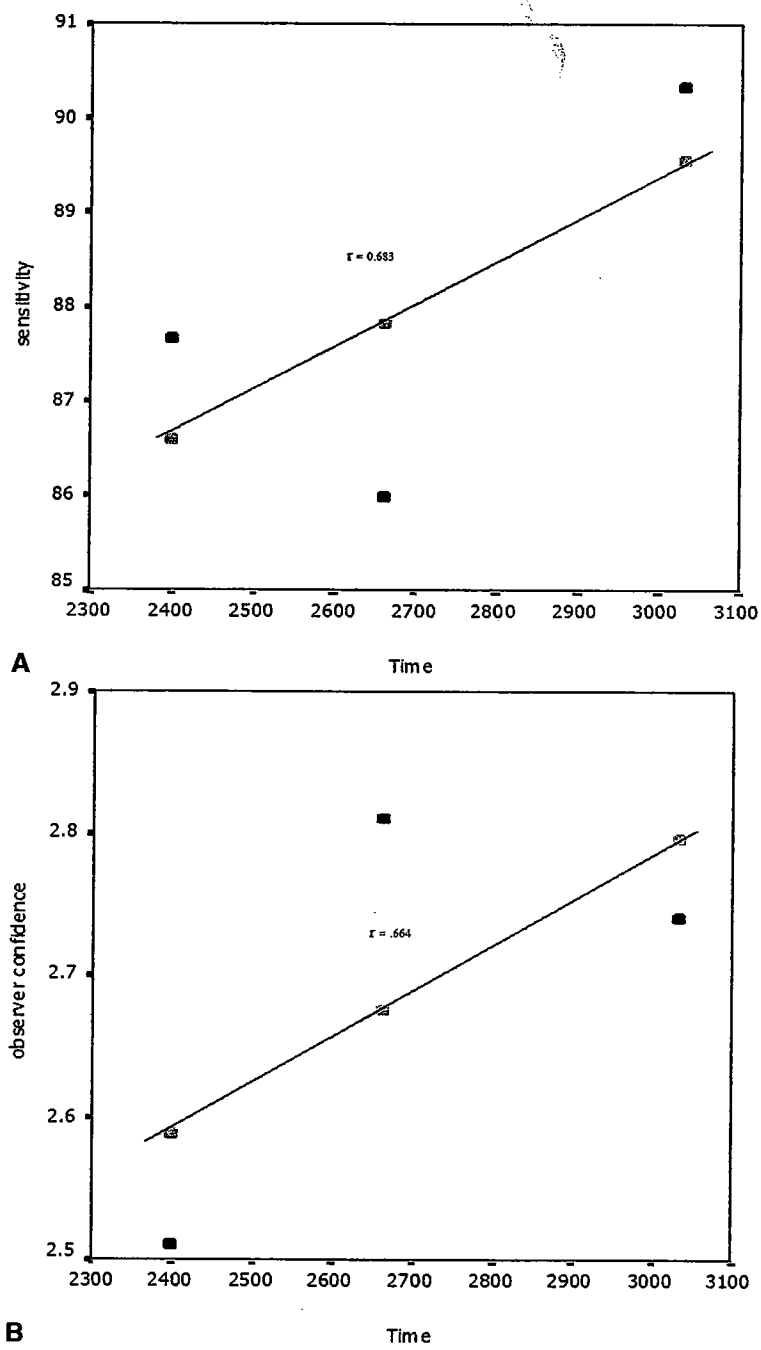


Fig 13. *A*: Time vs. sensitivity for Pattern 1 with evidence of strong correlation: $r = 0.63$; *B*: time vs. observer confidence for Pattern 1 with evidence of strong correlation: $r = 0.68$.

display configurations elicit similar observer response, although certain factors, such as lack of spatial noise in the background of each pattern, target difficulty, lack of time constraints, and the fixed viewing angle, each influence the results.

Although performing pair-wise comparisons drives up the type I error rate, we chose to do this rather than a three-way test in order to get additional data on how the displays relate to each other. Thus Bonfferroni correction had to be applied to reduced type I error rate.

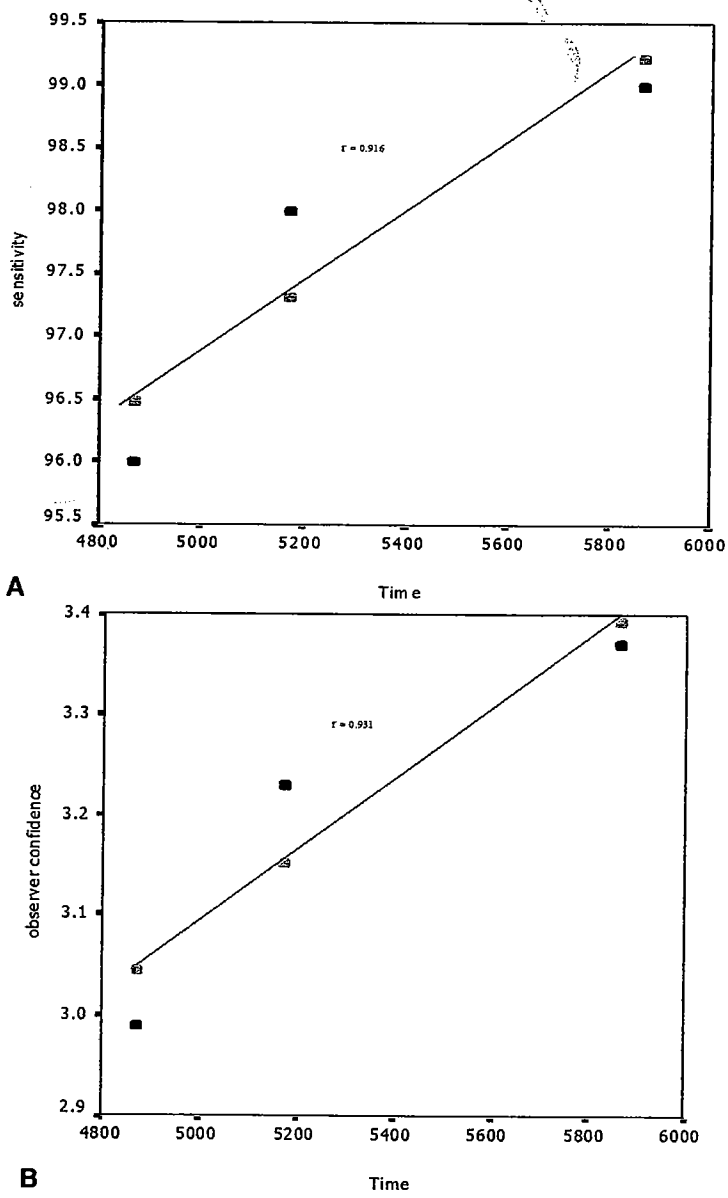


Fig 14. *A*: **Time vs. sensitivity for Pattern 2 with evidence of strong correlation: $r = 0.91$; *B*: time vs. observer Confidence for Pattern 2 with evidence of strong correlation: $r = 0.93$.**

We see a nearly 100% specificity as the observers had no trouble ruling out negative cases, which suggests to us that patterns with a noisier background would be more useful in predicting what outcome there may be in a clinical trail involving real medical images. Although the stimuli have covered a wide range of values in contrast/size, the high sensitivity values also suggest that the ratio of "difficult" to "easy" targets should possibly be increased as the former are more likely to be affected by monitor degradation issues. Thus a higher proportion of "difficult" targets is likely to demonstrate more differences between the displays. Introducing a time constraint could also have a negative affect on specificity. Time spent on each pattern had a strong correlation with sensitivity for all patterns, as well as with the observer confidence for patterns 1 and 2, which tested contrast-detail sensitivity. In the case of

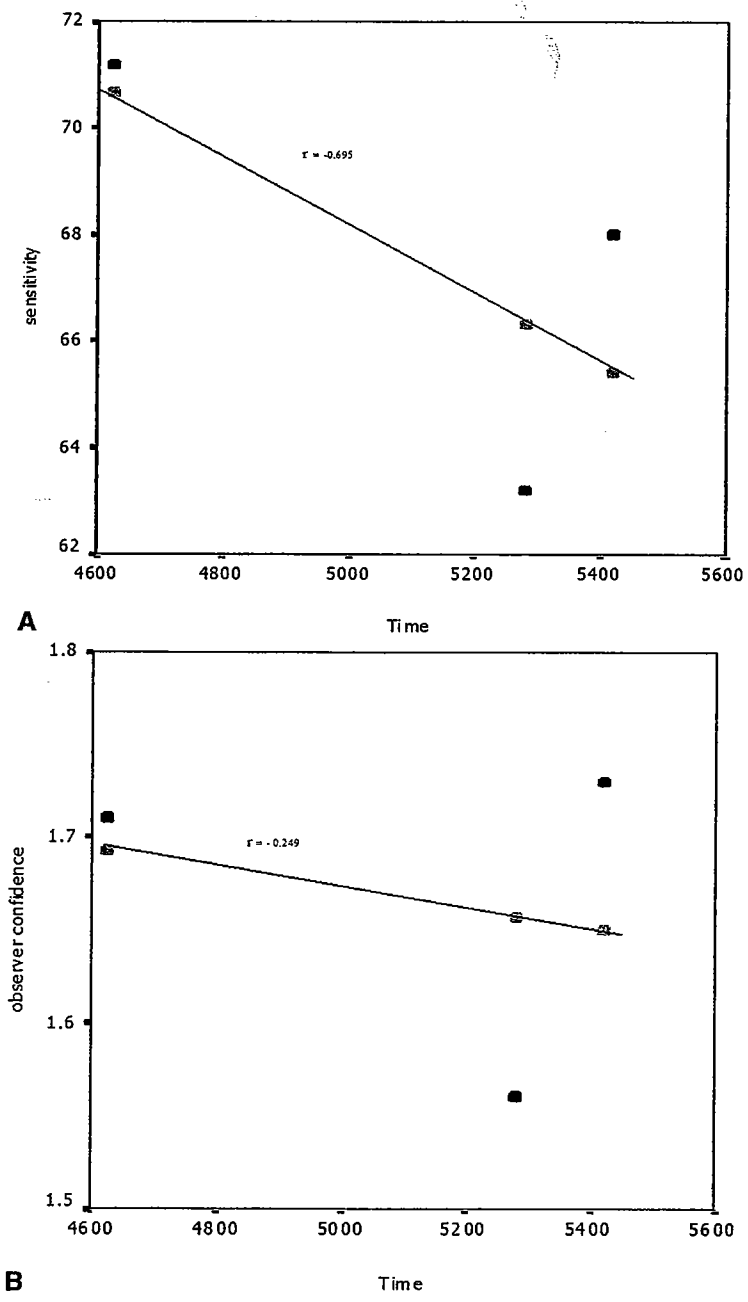


Fig 15. ***A*****: Time vs. sensitivity for Pattern 3 with evidence of negative correlation** $r = -0.68$**;** ***B*****: time vs observer confidence for Pattern 3. No significant correlation between time and observer confidence** $r = -0.25$**.**

pattern 3, the sensitivity actually decreased in relation to time spent, suggesting that for subtle patterns of this type, longer observation time did not provide for greater sensitivity. There was no significant correlation between observer confidence and time spent per pattern for patterns of type 3, which tested gray-scale grouping. This suggests to us that time did affect observer sensitivity for contrast-detail detection tasks, although it was not a significant factor in decision making for grouping tasks. Although there might be time constraints introduced by the physician schedules in an actual clinical setting, our feeling is that, in a trial situation,

the observers should be free to take the time necessary, as such constraints could significantly affect inter-observer variability.

The experimental set-up with the observers located a fixed distance away from each monitor provided consistency with respect to the subtended retinal viewing angle. Although this allowed us to compare the displays more objectively, the distance to any of the monitors was not ergonomically ideal, and in a clinical situation this distance would not be fixed. Thus we do not think that fixing this distance in a clinical trial is warranted.

CONCLUSION

Although an RGB calibrated IBM T221 monitor fared well in comparison to diagnostic quality displays, the results of this experiment do not answer the question of whether modern commercial LCDs can be used for medical image interpretation. A proper clinical trial must be conducted to unequivocally answer this question.

Acknowledgments

The authors thank Dr. Alfred Rademaker, Dr. Edward Hendrick, Dr. Daniel Ebroon, Dr. Arnon Macori, IBM, and Planar Technologies for their help and support.

REFERENCES

1. Flynn, MJ, Kanicki, J, Badano, A, et al: High-fidelity electronic display of digital radiographs. Radiographics 19:1653-1669, 1999

2. Roehrig H, Fan J, Furukawa T, et al.: Performance evaluation of LCD of displays. Proceedings CARS, pp 461-66, 2002

3. Blume H, Steven P, Cobb M, et al.: Characterization of high-resolution liquid crystal displays for medical images. Proc SPIE 4681:271-292, 2002

4. Wright SL, Samei E: Liquid crystal displays for medical imaging: a discussion of monochrome versus color. SPIE Medical Imaging; Visualization, Image-Guided Procedures and Display, Feb. 2004, SPIE Proceedings Vol. 5367, paper 49, 2004

5. Siegel E, Reiner B, Hooper F et al.: Clinical comparison of LCD and CRT. Society of Computer Applications in Radiology (SCAR) Gerber, Hogan, McAvoy, Mulligan, Qureshi, Sliker, Monitors in the Evaluation of Non-Displaced Fractures. Annual Meeting, Boston, June 2003

6. Yip, WM, Yim, WS, Knok, CS: ROC curve analysis of lesion detectability on phantoms : comparision of digital spot mammography with conventional spot mammography. Br J Radiol 74:621-628, 2001

7. Lewin, JM, D'Orsi, CJ, Hendrick, RE, et al: Clinical comparison of full-field digital mammography and screen-film mammography for detection of breast cancer. AJR Am J Roentgenol 179:671-677, 2002

8. Lewin, JM, Hendrick, RE, D'Orsi, CJ, et al: Comparison of full-field digital mammography with Screen-film mammography for cancer detection: results of 4,945, paired examinations. Radiology 218:873-880, 2001

9. Kuzmiak, CM, Millnamow, GA, Qaqish, B, et al: Comparison of full-field digital mammography with respect to dignostic accuracy of lesion characteristics in breast tissue biopsy specimens. Acade Radiol 12:1378-1382, 2002

10. Pisano, ED, Cole, EB, Kistner, EO, et al: Interpretation of digital mammograms: comparison of speed and accuracy of soft-copy versus printed-film display. Radiology 2:483-488, 2002

11. National Electronical Manufacturers Association: Digital Imaging and Communications in Medicine, DICOM part 14. http://medical.nema.org/dicom/2001/01_14PU.PDF

12. Averbukh, AN, Channin, DS, Flynn, MJ: Assessment of novel high resolution, color AMLCD for diagnostic medical image display: luminance performance and DICOM calibration. J of Digit Imaging 16:270-279, 2003

13. Flynn MF, Tchou P: Accurate measurement of monochrome luminance palettes for the calibration of medical LCD monitors. SPIE Medical Imaging; Visualization, Image Guided Procedures and Display, Feb. 2003, SPIE Proceedings Vol. 5029, paper 47, 2003

14. Acrum, DR: Viewing distance at computer workstations. Workplace Ergonomics .:10-12, 1996Sept./Oct

15. Ellis, WD: A Source Book of Gestalt Psychology New York: Harcourt, Brace & World, 1938

16. Wertheimer, M: Productive Thinking (Enlarged Ed.) New York: Harper & Row, 1989

17. Wertheimer M: Laws of organization in perceptual forms. First published as Untersuchungen zur Lehre von der Gestalt II, in Psycologische Forschung, 4, 301-350, 1923. Translation published in Ellis, W. (1938). A source book of Gestalt psychology (pp 71-88). London: Routledge & Kegan Paul, 1938

18. Miller, G: Simultaneous statistical inference2 New york: Springer Verlag, pages, 1981, pp 6-8

19. Keppel, G: Design and Analysis: A Researcher's Handbook Englewood Cliffs, NJ: Prentice-Hall, 1991