Journal of Digital Imaging

Medical Grade vs Off-the-Shelf Color Displays: Influence on Observer Performance and Visual Search

Elizabeth A. Krupinski

The goal of this study was to compare diagnostic accuracy of radiologists viewing clinical images on a top-of-the-line medical-grade vs a top-of-the-line commercial off-the-shelf (COTS) color display with the luminance values set to simulate a display that had been in use for 1 year. A set of 50 digital radiography chest images was selected for use in the study, half containing a solitary pulmonary nodule and half nodule-free. The images were displayed twice to each of six observers, once on each display. Eye position was recorded on a subset of the images. Overall, there was a statistically significant difference ($F = 4.1496$, $p = 0.0471$) between the medical-grade color display and the COTS color display in terms of receiver operating characteristic area under the curve values, with the medical-grade display yielding higher diagnostic accuracy. Total viewing time did not differ significantly, but eye position data revealed differences, suggesting better search and decision-making efficiency with the medical-grade display. Medical-grade color displays at 1 year old yield better diagnostic and search efficiency than COTS color displays and thus are recommended for primary reading if color displays are to be used.

KEY WORDS: Color displays, observer performance, eye tracking

INTRODUCTION

High-quality monochrome medical-grade displays are generally considered to be the standard for primary interpretation of most radiographic images. They have high resolution, high luminance, high contrast ratio, low noise, and a better modulation transfer function than most color displays. However, the technology used for color displays has changed[1] in recent years, and many radiology departments are either using or considering using color displays for primary reading. A number of recent studies[2–8] have examined diagnostic accuracy using color versus monochrome displays for radiologic image interpretation with the overall finding that many high-performance color displays yield diagnostic performance close to or equivalent to high-performance monochrome displays. The very important question that many departments are asking today, however, is whether commercial off-the-shelf (COTS) color displays are adequate or are high-performance medical-grade color displays required.

One question in particular is how well color displays perform as they age. It is well known that there are visible changes as monochrome displays age, reducing their utility over time[9]. It has also been shown that as monochrome liquid crystal displays (LCDs) age, there are slight degradations in diagnostic performance and visual search efficiency[10]. Radiologists tend to take a little longer to render a positive decision and tend to fixate missed lesions for less time with older displays. Whether the same types of performance degradations occur with medical-grade versus COTS color LCDs has yet to be explored. Thus, the goal of this study was to compare decision and

From the Department of Radiology, University of Arizona, 1609 N. Warren Bldg 211 Rm 112, Tucson, AZ 85724, USA.

Correspondence to: Elizabeth A. Krupinski Department of Radiology, University of Arizona, 1609 N. Warren Bldg 211 Rm 112, Tucson, AZ 85724, USA; tel: +1-520-6264498; fax: +1-520-6264376; e-mail: Krupinski@radiology.arizona.edu

Online publication 3 September 2008
doi: 10.1007/s10278-008-9156-6

visual search performance using a medical-grade color LCD and a COTS color display that have been set to luminance levels that would be expected after 1 year of use. Medical-grade displays often come with technology that automatically monitors luminance levels and adjusts the display luminance back to original levels as they drift over time. Few if any COTS displays come with this type of technology, and thus it is up to the user to monitor and adjust the display luminance as it drifts over time. In a typical clinical setting, this is not likely to be a problem since monitors tend to be checked regularly and adjusted by the picture archiving and communication systems manager and technology team. When radiologists are at home reading cases or other nonstandard environments, it is less likely that they would have a luminance meter to measure and adjust display luminance on a regular basis, and thus luminance drifts may not be adjusted for as they should.

MATERIALS AND METHODS

The study consisted of two parts—an observer study and an eye position study, both using the same observers. A set of 50 digital radiography chest images (posteroanterior view) was selected for use in the study. Half of the images contained a solitary pulmonary nodule verified by computed tomography (CT), and half were nodule-free as verified by CT. The nodules ranged in size from 9 to 26 mm and were rated as round or round with some speculations and as subtle to moderately subtle by the experienced chest radiologist not serving as a reader. The images were displayed twice to each of six observers (three senior residents and three board-certified radiologists). In one trial, the images were displayed on the 3-MP Barco color medical-grade display (Barco Coronis MDCC-3120-DL), and in the other, they were displayed on the COTS color 2-MP monitor (Dell 2405). Each display was calibrated to luminance that corresponds to backlight aging after 1-year time and were calibrated to the Digital Imaging and Communications in Medicine Grayscale Standard Display Function. The maximum luminance on the Barco display was 500 cd/m^2 with a minimum of 0.77 cd/m^2. The maximum luminance on the Dell COTS was 342 cd/m^2 with a minimum of 0.376 cd/m^2. Trials were separated by a minimum of 3 weeks to promote forgetting. Half of the participants used the Barco display first, and half used the COTS display first. Images were displayed using the custom-developed ImprocRAD (v. 5.52) image display software[11]. For each trial, the observer viewed the image and then reported their decision using a six-point scale. They reported whether or not a nodule was present and then their confidence in that decision as possible, probable or definite. Each image trial was timed beginning with the point at which the image appeared on the display and ending when the observer selected the next image button.

A subset of 20 of the images (five nodule-free, 15 nodule-containing) was then used in an eye position study with the six observers where eye position was recorded. They observers returned after at least 1 month from their second session in the first study to promote forgetting. Memory effects are unlikely to be avoided (even after many months), but about 1 month of daily reading of clinical cases is usually sufficient to render most study cases forgettable. The eye-tracking study was carried out separately from the observer study for two reasons. The first is that eye position studies typically do not require as many images as an receiver operating characteristic (ROC) study so this part used only 20 instead of the full 50. The second reason is that although the eye tracker is not uncomfortable, after about 30 min, observers do tend to start to feel its weight so we try to limit the length of these studies to about 30 min total.

The observers viewed the images twice, once on the Barco medical-grade display and once on the COTS display. A counterbalanced design was used and at least 3 weeks passed between sessions. The study used the 4000SU Eye-Tracker with head tracker (Applied Science Labs, Bedford, MA, USA). This infrared-based system uses pupil and corneal reflections to track the line of gaze to within 1° of accuracy. The head tracker allows the observer to move once the system has been calibrated[12]. The SU4000 samples eye position every 1/60 s to generate x,y coordinate data indicating the location of gaze. Using a running mean distance and temporal threshold algorithm, the raw x,y data calculates fixations and the dwell time associated with them. These fixations can be correlated with image locations and the decisions rendered about the specific locations. The eye

position data[12] was used to characterize the following: time to first fixate a lesion, total search time, and dwells associated with each decision type (true and false, positive and negative). These are the parameters we have analyzed in a number of other teleradiology and telepathology studies[13–16].

RESULTS

Diagnostic Accuracy

The confidence data were analyzed using the multireader multicase ROC analysis of Dorfman et al. to generate ROC area under the curve (Az) values and compare them statistically[17]. Overall, there was a statistically significant difference (F=4.1496, p=0.0471; 95% confidence interval=0.0009, 0.1346) between the medical-grade color display (mean ROC Az=0.9101, SE=0.0193) and the COTS color display (mean ROC Az=0.8424, SE=0.0169). The individual and mean ROC Az values are shown in Figure 1. The overall sensitivity with the medical-grade display was 0.91 with a specificity of 0.93. The overall sensitivity with the COTS display was 0.86 with a specificity of 0.92.

Although average viewing time was longer with the COTS display (mean=37.99 s, SD=18.36; range=7–68 s) than with the medical-grade display (mean=35.17 s, SD=17.73; range=6–65 s), the difference was not statistically significant (F=3.38, p=0.067). The individual and mean viewing times are shown in Figure 2.

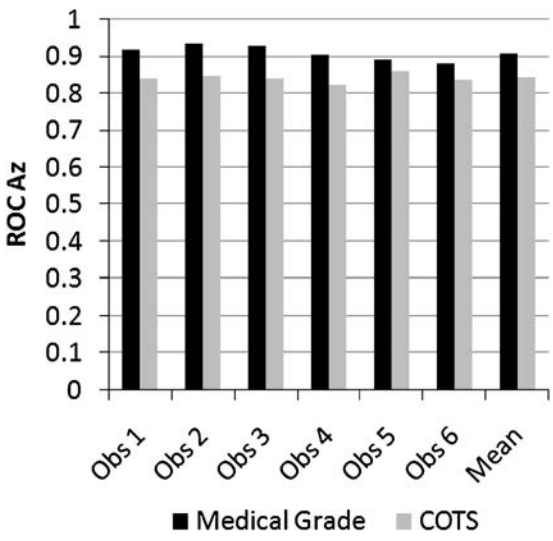

Fig 1. Individual and mean ROC Az values for the medical-grade vs COTS color displays. Observers 1, 2, and 3 (*left*) are the radiologists, and observers 4, 5, and 6 (*right*) are the residents.

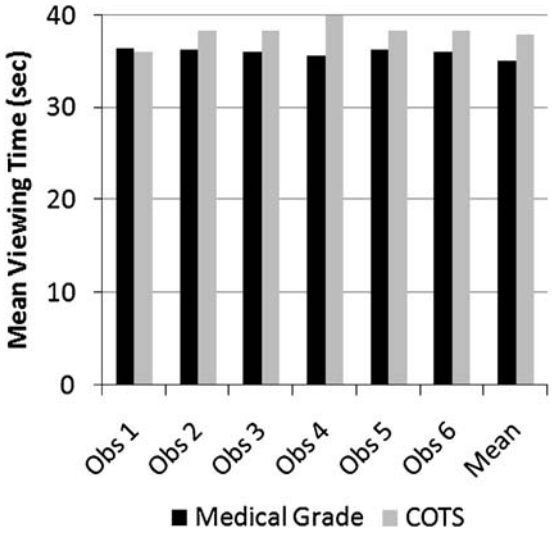

Fig 2. Individual and mean viewing times for medical-grade vs COTS color displays. Observers 1, 2, and 3 (*left*) are the radiologists, and observers 4, 5, and 6 (*right*) are the residents.

Eye Position Analyses

In the eye position-recording study, there was no statistically significant difference in total viewing time as a function of display (F=1.064, p=0.3034) or whether the image contained a nodule (F=0.215, p=0.6435). Overall viewing times were, however, slightly shorter with the medical-grade display than with the COTS display (see Fig. 3).

The second parameter measured was the time to first hit the nodule (true positive [TP] or false negative [FN] report) or non-nodule (false positive [FP]) reported location (see Fig. 4). The time to first hit is defined as the length of time into search, beginning at the point when the image appears and the eye position recording has begun until a fixation first lands on the nodule or FP location. For the TP reports, there was no statistically significant difference (t=1.398, p=0.164) in time to first hit the nodule for medical-grade (mean=1.531 s, SD=1.483, n=86) vs COTS (mean=1.30 s, SD=0.823, n=84) displays. There was no difference (t=0.337, p=0.745) for medical grade (mean=1.622 s, SD=0.290, n=4) vs COTS

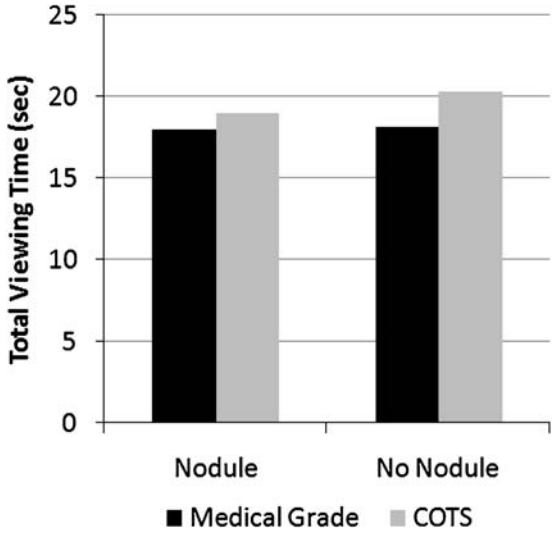

Fig 3. Mean total viewing times for medical-grade vs COTS color displays in the eye position study.

(mean=1.87 s, SD=1.911, n=6) for FNs. The difference between medical grade (mean=3.987 s, SD=1.759, n=3) and COTS (mean=2.282 s, SD=0.525, n=5) for FP reports approached but did not reach statistical significance (t=2.413, p=0.0524).

The final parameter measured was cumulative dwell time associated with each type of decision (see Fig. 5). Cumulative dwell time per decision is the sum of all fixation clusters landing within 2.5° of a nodule (whether or not it is reported, TP and FN decisions or a location falsely indicated as a nodule [FP]). All fixation clusters not associated with a true or false nodule location are considered true negative (TN) decisions. Although cumulative dwell times for TP decisions were longer for the COTS (mean=7.077 s, SD=22.916) than the medical-grade display (mean=5.913 s, SD=14.577), the difference did not reach statistical significance (t=1.759, p=0.0803). Cumulative dwell times for FP decisions were also longer for the COTS (mean=2.542 s, SD=2.265) than medical-grade (mean=1.24 s, SD=0.618) displays but again did not reach significance (t=1.361, p=0.2223). Cumulative dwell times for the FN decisions were shorter for the COTS (mean=3.117 s, SD=4.035) than medical-grade displays (mean=4.292 s, SD=5.196) but did not reach statistical significance (t=0.818, p=0.4369). There was a statistically significant difference (t=3.136, p=0.0017) between COTS (mean=0.367 s, SD=0.135, n=4,623) and medical-grade (mean=0.344 s, SD=0.115, n=4,759) displays for TN decisions.

DISCUSSION

Overall, diagnostic accuracy was significantly higher with the medical-grade than COTS color displays in terms of ROC Az and sensitivity and specificity. The nodules were rated by an indepen-

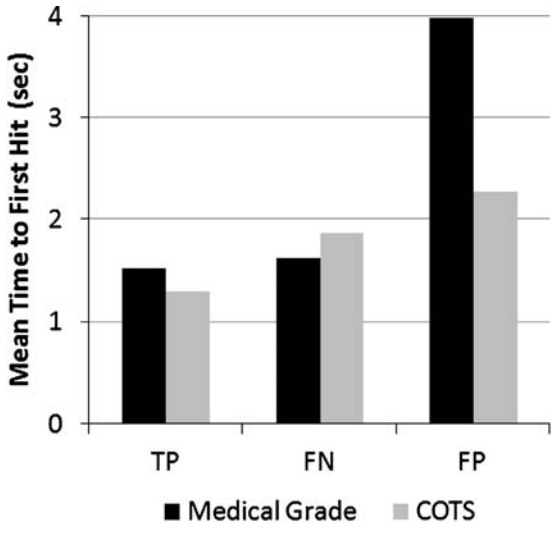

Fig 4. Mean time to first fixate the nodules (*TP* or *FN* decision) and non-nodule (*FP*) reported locations for medical-grade vs COTS color displays.

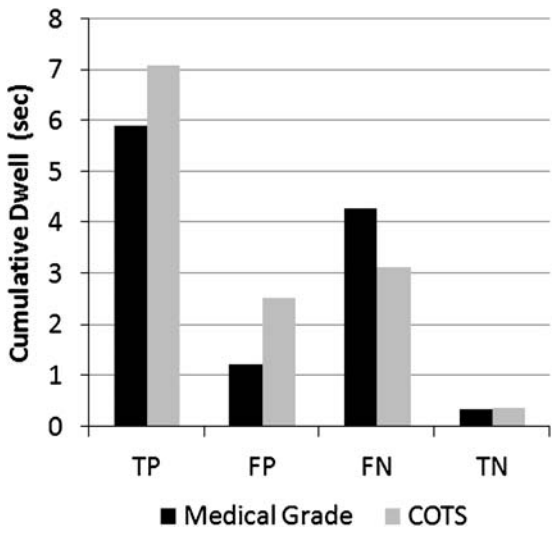

Fig 5. Mean cumulative dwell time for each decision (*TP, FP, FN, TN*) for medical-grade vs COTS color displays.

dent radiologist as ranging from subtle to moderately subtle, so although performance was relatively high, these are typical clinical cases. It is possible that the use of a test set with only nodules in it with the task specifically to search only for nodules does improve performance as compared to being in the clinic with images that contain a host of lesion types with and without clinical history. There has actually been little or no study on this particular phenomenon, although there is always the general concern that laboratory studies do not exactly replicate the clinical setting. This is true, but if the test set was influencing performance, it would have done so equally for both display types in all likelihood, so any differences that we observed were in all likelihood attributable to the displays themselves.

The eye position data suggest that reading efficiency is better with the medical-grade than COTS displays. The eye position data revealed that total viewing time was slightly longer with the COTS than medical-grade display as expected. Cumulative dwell times associated with the TP and FP decisions on the medical-grade display were shorter than with either of the COTS display, suggesting that it took less visual processing or attentional effort to discriminate the nodule from the background with this display. The observers spent longer looking at the nodules when they did not report them (FNs), suggesting that there was more information or features that were attracting attention with the medical-grade display compared to the COTS display. Cumulative dwell times for TN decisions were statistically shorter with the medical-grade than COTS displays but only by 0.023 s. The reason it reached statistical significance is because of the very high number of TN decisions made during search—all image locations viewed that are not associated with a TP, FN, or FP decision are TN decisions by default. There were 4,759 TN decisions on the medical-grade display overall and 4,623 on the COTS display overall. Although the difference reached statistical significance, it seems unlikely that such a small difference would have practical significance. Although it could be a factor influencing total reading times, further studies need to be done. The time to first fixate the nodules revealed no statistically significant differences between the two displays.

CONCLUSIONS

Although COTS color displays may be less expensive to purchase compared to medical-grade color displays, the medical-grade displays are generally more stable over time in terms of maintaining consistent levels of backlighting and thus luminance levels. The results of this study suggest that after just 1 year of use, the COTS display may degrade enough to negatively impact diagnostic and visual search performance. Clearly, COTS displays are continually improving, but care should still be taken when considering whether to purchase medical-grade versus COTS color displays for primary diagnostic interpretation. At the very least, color monitors of any type need to be calibrated and evaluated on a regular (at least once every 6 months for the first 2 years and every 4 months thereafter) basis[18].

ACKNOWLEDGMENTS

This work was supported in part by Barco.

REFERENCES

1. Barco: Color your world: diagnostic color displays take off. Health Imaging IT 4:30–33, 2006

2. Doyle AJ, LeFevre J, Anderson GD: Personal computer versus workstation display: observer performance in detection of wrist fractures on digital radiographs. Radiol 237:872–877, 2005

3. Lehmkuhl L, Mulzer J, Teichgraeber U, Gillessen C, Ehrenstein T, Ricke J: Evaluation of the display quality of different modalities in digital radiology. Fortschr Geb Rontgenstr Nuklearmed 176:1031–1038, 2004

4. Ricke J, Hanninen EL, Zielinski C, Amthauer H, Stroszczynski C, Liebig T, Wolf M, Hosten N: Shortcomings of low-cost imaging systems for viewing computed radiographs. Comput Med Imaging Graph 24:25–32, 2000

5. Langer S, Fetterly K, Mandrekar J, Harmsen S, Bartholmai B, Patton C, Bishop A, McCannel C: ROC study of four LCD displays under typical medical center lighting conditions. J Dig Imag 19:30–40, 2006

6. Averbukh AN, Channin DS, Homhual P: Comparison of human observer performance of contrast-detail detection across multiple liquid crystal displays. J Digit Imaging 18:66–77, 2005

7. Hirschorn D: Consumer displays for radiography? You've got to be kidding. SIIM News 19:5/19, 2007

8. Krupinski EA, Roehrig H, Fan J, Yoneda T: Monochrome versus color softcopy displays for teleradiology: observer performance and visual search efficiency. Telemed e-Health 13:675–681, 2007

9. Roehrig H: The monochrome cathode ray tube display and its performance. In: Kim Y, Horii SC Eds. Handbook of

Medical Imaging: Display and PACS. volume 3, Bellingham, WA: SPIE Press, 2000, pp. 155–220

10. Krupinski EA, Roehrig H, Fan J: Does the age of liquid crystal displays influence observer performance? Acad Radiol 14:463–467, 2007

11. Dallas WJ, Roehrig H: ImprocRAD software components for mammogram display and analysis. Proc SPIE Med Imag 4322:1129–1140, 2001

12. Nodine CF, Kundel HL, Toto LC, Krupinski EA: Recording and analyzing eye-position data using a microcomputer workstation. Behav Res Methods Instrum Comput 24:475–485, 1992

13. Krupinski EA: Visual scanning patterns of radiologists searching mammograms. Acad Radiol 3:137–144, 1996

14. Krupinski EA: Visual search of mammographic images: influence of lesion subtlety. Acad Radiol 12:965–969, 2005

15. Krupinski EA, Roehrig H: The influence of perceptually linearized display on observer performance and visual search. Acad Radiol 7:8–13, 2000

16. Krupinski EA, Tillack AA, Richter L, Henderson JT, Bhattacharyya AK, Scott KM, Graham AR, Descour MR, Davis JR, Weinstein RS: Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. Human Pathol 37:1543–1556, 2006

17. Dorfman DD, Berbaum KS, Metz CE: Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol 27:723–731, 1996

18. DICOM. http://medical.nema.org/. Last accessed March 27, 2008.