Journal of Digital Imaging

ROC Study of Four LCD Displays Under Typical Medical Center Lighting Conditions

Steve Langer, Ken Fetterly, Jay Mandrekar, Scott Harmsen, Brian Bartholmai, Charles Patton, Alan Bishop, and Colin McCannel

Nine observers reviewed a previously assembled library of 320 chest computed radiography (CR) images. Observers participated in four sessions, reading a different 1/4 of the sample on each of four liquid crystal displays: a 2-megapixel (MP) consumer color display, a 2-MP business color display, a 2-MP medical-grade gray display, and a 3-MP gray display. Each display was calibrated according to the DICOM Part 14 standard. The viewing application required observer login, then randomized the order of the subsample seen on the display, and timed the responses of the observer to render a 1–5 judgment on the absence or presence of ILD on chest CRs. Selections of 1–2 were considered negative, 3 was indeterminate, and 4–5 were positive. The order of viewing sessions was also randomized for each observer. The experiment was conducted under controlled lighting, temperature, and sound conditions to mimic conditions typically found in a patient examination room. Lighting was indirect, and illuminance at the display face was 195 ± 8% lux and was monitored over the course of the experiment. The average observer sensitivity for the 2 MP color consumer, 2 MP business color, 2 MP gray, and 3 MP gray displays were 83.7%, 84.1%, 85.5%, and 86.7%, respectively. The only pairwise significant difference was between the 2-MP consumer color and the 2-MP gray ($P = 0.05$). Effect of order within a session was not significant ($P = 0.21$): period 1 (84.3%), period 2 (86.2%), period 3 (85.4%), period 4 (84.1%). Observer specificity for the various displays was not statistically significant ($P = 0.21$). Finally, a timing analysis showed no significant difference between the displays for the user group ($P = 0.13$), ranging from 5.3 s (2 MP color business) to 5.9 s (3 MP Gray). There was, however, a reduction in time over the study that was significant ($P < 0.001$) for all users; the group average decreased from 6.5 to 4.7 s per image. Physical measurements of the resolution, contrast, and noise properties of the displays were acquired. Most notably, the noise of the displays varied by 3.5× between the lowest and highest noise displays. Differences in display noise were indicative of observer performance. However, the large difference in the magnitude of the noise was not predictive of the small difference (3%) in the observer sensitivity for various displays. This is likely because detection of interstitial lung disease is limited by "anatomical noise" rather than display or x-ray image noise.

KEY WORDS: ROC, image quality, displays, interstitial lung disease

INTRODUCTION

The standard of care for clinical image review is highly variable: from Web-based image viewers on SVGA (1024 × 768) displays to thick clients on dedicated 2- or 3-megapixel (MP) medical grade displays. The corresponding cost per display is also highly variable: $300–7000 or more per display. The display performance requirements also vary considerably depending upon the area the clinical workstation is used in. Workstations in emergency rooms (ERs) and intensive care units (ICUs) may be called upon to provide critical care decisions in advance of the radiologist's report rendered on a picture archival and communication system (PACS), whereas displays in exam rooms may be largely used for patient education.

Our institution considered all these variables in the search for a clinical viewing platform. In the end, it was decided to advance two classes of clinical viewers and keep the PACS-connected

From the Mayo Clinic and Foundation, Department of Radiology, 200 First Street SW, Rochester, MN 55905, USA.

Correspondence to: Steve Langer, Mayo Clinic and Foundation, Department of Radiology, 200 First Street SW, Rochester, MN 55905, USA, tel: +507-266-4418; e-mail: langer.steve@mayo.edu

Online publication 28 October 2005
doi: 10.1007/s10278-005-8149-y

workstations in the Radiology department. The base class of clinical viewer, of which we have 15,000 units, would remain at an SVGA class display. The more demanding clinical viewer, of which we estimate there will be 500–1500 units, was deemed to require a dual-display viewing application. The question—what type of display?

MATERIALS AND METHODS

Image Acquisition, Processing and Display Platform

All images used in the study were acquired from Fuji computed radiography (CR) systems and archived without persistent application of any vendor-specific lookup tables (LUTs), edge enhancement algorithms, or annotations. The images were stored on institutionally developed display software with the Window/Level optimized for default appearance, to minimize the need for radiologist observers to manipulate or process the images. The images were displayed to fit the screen using a bicubic downsampling algorithm. The selected displays were:

- 2 MP Dell "the 2 MP color consumer class" (Dell Inc., Round Rock, TX, USA)
- 2 MP NEC 2180 "2 MP color professional class" (NEC, New York, NY, USA)
- 2 MP Siemens "2 MP grayscale medical grade" (Siemens AG, Munich, Germany)
- 3 MP Siemens "3 MP grayscale medical grade" (Siemens AG, Munich, Germany)

The upper limit of a 3-MP grayscale medical display was experimentally derived from a previous work.[8]

Sample Selection

Computed radiographs of the chest were utilized in this study to maximize the potential to discover differences in display capabilities. In particular, the low-contrast, high-spatial-frequency abnormalities present in interstitial lung disease (ILD) potentially allow for differentiation in diagnostic sensitivity between display devices of different matrix size, contrast properties, and noise properties.

Internal Review Board (IRB) approval was obtained to perform a retrospective record search and to utilize anonymized image data for the study. Specifically, a retrospective review of a radiology database that tracks diagnoses on CT scans was performed for all studies from 1997 to 2003. This search specifically targeted chest CT scans coded for ILD. From the results of this search, a secondary correlative search of the radiology information system was performed to extract a subset of CT-proven cases of ILD that also had CR chest radiographs performed within 6 months of the CT scan. The CT studies discovered as a result of these searches were each pulled from the radiology archive and were reviewed by a board-certified radiologist specializing in Thoracic Radiology (B.J.B.). This radiologist confirmed the presence, ascertained the type, and determined the extent of ILD on each of these cases. For the purposes of this study, studies with classic features of usual interstitial pneumonitis (UIP), nonspecific interstitial pneumonitis (NSIP), chronic hypersensitivity pneumonitis (HSP), and other diffuse atypical fibrotic lung disease were included. Focal fibrotic changes, diffuse alveolar processes, or studies with predominantly ground-glass opacities were excluded from the study. In addition, where possible, cases of mild or early manifestations of ILD were preferred for inclusion—to optimize the possibility of detecting differences in the two types of displays utilized in the study. Studies with significant pathology unrelated to ILD were also excluded from the study.

Each of the CR images for CT-proven ILD patients were also reviewed by a board-certified Thoracic Radiology specialist (B.J.B.). As with the selection of CT examinations, CR images with the most subtle manifestations of disease were favored for inclusion in the study, and any images with significant pathology unrelated to the proven ILD were excluded. For some patients, multiple CR studies obtained on different dates were utilized. A similar method was utilized to obtain normal CR images for inclusion in the study. A search of the radiology diagnostic database over the course of 1 year was used to discover negative CTs of the chest. For each of these results, a correlative search for chest CR examinations within 6 months of the CT was performed in the radiology information system. For each of the CR studies found, the images were evaluated by a board-certified Thoracic Radiology specialist and any images with pathology or transient abnormalities such as atelectasis were excluded. For the study, the individual normal and abnormal images were deidentified, pooled, and randomized.

Data Collection

Nine observers were assembled: four from Thoracic Radiology and five from Pulmonology. The observers participated in four sessions, and in each session they read a different 1/4 of the sample on each of four displays: a 2-MP consumer color display, a 2-MP professional color display, a 2-MP medical-grade gray display, and a 3-MP gray display. The viewing application required each observer to log in, then randomized the order of the subsample seen on the display, and timed the responses of the observer to render a 1–5 judgment on the absence or presence of ILD on the chest CR. Selections of 1–2 were considered negative, 3 was indeterminate, and 4–5 were positive. The order of images was also randomized for each observer session so that an observer never saw the same images in the same order as they viewed a given sample across all four displays. At the end of four sessions, each observer had seen each sample quartile on each display, and the image randomization disrupted any learning effect (as will be seen). Observers could not adjust the images in any manner. Specific ambient lighting, temperature, and sound conditions were used to mimic conditions typically found in examination rooms. Lighting was indirect, and illuminance at the display face was 195 ± 8% lux and was monitored over the entire course of the experiment.

Data Analysis

The study included a total of 310 images: 230 CT-proven positive chest images, and another 80 normal chest images were included to maintain reader integrity. The images were randomly assigned to four reading sets. Stratified randomization was used with 20 normal images in each of the four sets and the remaining 230 positive images assigned across the sets. The sample set provides an 80% power to detect a significant difference between any two display's sensitivities of 80% vs. less than 70% assuming a one-sided McNemar's test of paired proportions at an alpha level of 0.05.[5,9]

The experimental design included five factors: display type, observer, day, order in day, and image sets. The five level observer response (1 = negative, 5 = positive) was dichotomized as positive if 3, 4, or 5, and negative if 1 or 2. For each of the observer–display–image sets, estimates of sensitivity and specificity for positive chest abnormality were calculated; sensitivity estimates were treated as the response variable in an analysis of variance. The interaction was included to allow for different display performances across different observers. Pairwise significance tests within each of the five main effect terms in this model were tested using the Tukey–Kramer multiple comparison procedure.[6,7,10] Linear, quadratic, and cubic relationships of day and order in day with sensitivity estimates were also tested. Specificity was analyzed similarly to sensitivity. In addition, the average viewing time (the time from when an image appears on the display until an observer enters their decision) was calculated for each of the observer–display–image sets. These average viewing times were used as the response in an analysis of variance, including the same five factors used in the analysis of sensitivity.

Display Setup, Calibration and Measurements

Display Setup

All of the displays were calibrated to the Digital Imaging and Communications in Medicine (DICOM) Part 14: Grayscale Standard Display Function.[3] The vendor-specified maximum luminance of the 2-MP color displays was 250 cd/m^2 and that of the 2- and 3-MP grayscale displays was 600 cd/m^2. In practice, it is expected that the maximum luminance of a display can be achieved for a relatively short period of time because of backlight output decay. To mimic a practical clinical implementation, the maximum luminance of the calibrated displays was set to 200 cd/m^2 for the color displays and 400 cd/m^2 for the grayscale displays. An independent company (Image Smiths Verilum, Germantown, MD, USA) performed calibration of the color displays. Both of the color displays used were provided to Image Smiths and used to create unit-specific calibrations. Image Smiths used the OptiGrayscale routine of the Verilum software package to calibrate the color displays. Each of the three subpixels of a display pixel can be separately addressed with an 8-bit digital signal. The OptiGrayscale calibration method uses subpixel modulation to produce a grayscale look-up table containing approximately 10^3 unique gray levels. A grayscale LUT value is assigned for each of the 256 points of the 8-bit display input range such that the luminance output of the display matches (as close as possible) the luminance output indicated by the DGSDF. As these are color displays, subpixel modulation might be expected to affect the color tone of the gray levels. The OptiGrayscale routine recognizes this and does not allow interpolated gray levels that have perceivably different color tone. Examination of images and test patterns presented on the color displays thus calibrated did not demonstrate perceivable color differences in the displayed gray levels. The grayscale displays were calibrated using the manufacturer provided calibration software (SMFitACT 3.0; Siemens Displays, Germany) and photometer (SMFitACT Spotmeter; Wellhofer, Germany). This software also uses subpixel modulation to increase the effective number of possible LUT gray levels. The minimum luminance of the displays was not specifically set, but rather left at the "native" minimum luminance of the display.

Display Characterization

The luminance output of each display was measured for each pixel value over the 8-bit input range. A scientific-grade CCD camera (Orca ER; Hamamatsu, Japan) with a photographic lens (Micro-Nikkor 105 mm; Nikon, Japan) was used as the photometer. The pixel values of the camera are linearly proportional to input light fluence. The linear response function of the camera was uniquely characterized each time the camera was directed upon a display. For camera calibration, the luminance response of the display was first measured at 18 uniformly spaced digital driving levels (DDLs) over the 8-bit grayscale range using a calibrated UDT 371 meter with a model #265 luminance probe (UDT Instruments, Baltimore, MD, USA). The camera was positioned and focused, and then photographic images of the 18-step images were acquired. The average value of the image pixel values (PV) was calculated to create (DDL, $\overline{\mathrm{PV}}$) data pairs. DDL values resulting in luminance inputs that exceeded the dynamic range of the camera (with fixed exposure time) were manually identified and discarded, leaving at least 15 useful data pairs for each display. When used as a photometer, the exposure time of the camera was programmatically adjusted to ensure that the light fluence incident upon the CCD was within the linear range of the camera. The luminance was calculated from the photographic images using the transform

$$L = \frac{t_0}{t}\,\frac{\overline{\mathrm{PV}} - I}{S}, \tag{1}$$

where t_0 is the exposure time used for the camera calibration, t is the exposure time corresponding to the photometric image acquired, $\overline{\mathrm{PV}}$ is the average image pixel value of the image, I is the linear fit intercept, and S is the linear fit slope. Luminance measurements acquired with the spot photometer and the camera agreed to within 2%.

The luminance at each of the 256 steps of the 8-bit displays was measured using the calibrated camera as a photometer. The measured luminance values were converted to perceivable just noticeable difference (JND) index values using a combination logarithmic–polynomial equation described in the DGSDF. Next, the difference between the JND indices

measured at adjacent DDL values was divided by the difference between the adjacent DDLs (1, in this case), resulting in a measurement of perceivable contrast as ΔJND/ΔDDL. An ideal display system would have equal ΔJND/ΔDDL values across the entire 8-bit display range. Deviation from this ideal function was characterized as the root mean square error between the average of the ΔJND/ΔDDL values and the individual ΔJND/ΔDDL values.

Measurements of display luminance were acquired periodically (nominally every 10 days) to ensure consistent display performance throughout the course of the experiment. The luminance was measured at five equally spaced DDL values across the 8-bit range, and the total JND range of each display was calculated. The ambient light incident upon the face of the displays was also measured periodically using the UDT 371 meter and cosine diffuser probe.

Preliminary measurements of the diffuse reflection properties of the display were acquired. The displays were placed in a 10×10 ft^2 room with uniform wall coverings and diffuse fluorescent lighting. The wall coverings and lighting were consistent with those of a typical office or examination room and similar to those in which the observer experiments were conducted. The room illumination was varied over the range 40–200 lux and the light reflected from the face of the displays was measured with the narrow-angle photometer. The coefficient of diffuse reflection (R_d) of the displays was calculated as

$$R_d = \frac{L_r}{I}, \quad (2)$$

where I is the illuminance (lux) incident upon the display and L_r is the reflected luminance (cd/m^2).

The modulation transfer function (MTF) has been used by other authors to specify the spatial resolution properties of electronic displays.[1,2] Preliminary measurements of the MTF of the displays were acquired using the digital camera described above. The MTF measurements demonstrated that the spatial blur of these LCDs was very small. This conclusion is consistent with the findings of other authors;[2] therefore, MTF measurement results will not be presented here.

Visual examination of the displays demonstrated that each display presented a varying level of perceived texture, or visual noise. The AAPM TG-18 report recommends quantitative assessment of display noise through measurement of the normalized noise power spectrum (nNPS). The nNPS provides a measurement of display noise as a function of spatial frequency. Although the NPS may be directly measured, it is of limited practical use in and of itself. Of greater relevance is a comparison of the noise properties of the displays compared to that of other sources of noise in an x-ray imaging chain. Experiments were conducted to determine the magnitude of the display nNPS relative to that of computed radiography (CR) digital x-ray images. The simplistic methods (and results) presented here may be considered to be preliminary in that they do not fully represent the complex nature of noise in x-ray imaging and image visualization. However, the noise inherent in the x-ray images and that of the display are the two major sources of physical noise in the imaging chain and are worthy of at least some consideration.

The purpose of the nNPS measurements was to compare the magnitude of the display noise to that of a typical CR x-ray image. A single uniform x-ray image was acquired using an exposure of 1 milli-Roentgen (mR) from a diagnostically relevant x-ray beam. The pixel pitch of this image was 0.1 mm and it was displayed by using a 1:1 image pixel to display pixel mapping. A 1344×1024 photograph of a small region of the CR noise image was acquired. The pixel pitch of the photographic images, defined at the image plane, was 0.025 mm. The nNPS was calculated in a manner similar to those used by other authors[1,2] and to methods used for evaluating digital x-ray imaging devices.[4] At this point, the nNPS contained noise contributions from the quantum uncertainty of the light fluence incident upon the camera and small contributions from CCD camera noise. These contributions were characterized and subsequently removed from the nNPS. Next, the nNPS was corrected for camera blur by dividing by the square of the camera MTF. The resultant nNPS contained noise contributions from the x-ray image and the display. The 10-bit CR image had an average gray level of 445, which scales to a digital driving level of 111 in the 8-bit display range. A uniform digital test pattern with DDL = 111 was displayed and photographed. Note that the camera was not moved between photographs, resulting in exactly the same region of the display sampled for the two images. The nNPS of the display only was calculated in the same manner as described above for the CR and display nNPS. The difference between the nNPS acquired with the CR and display and the display alone was calculated. This difference is expected to represent the propagation of the CR noise through the display system. Note that the nNPS was calculated in 2D, then reduced to 1D for presentation. The nNPS presented in Results and Discussion represents the average of the orthogonal 1D axial measurements.

RESULTS AND DISCUSSION

Observer Results

Time Analysis Per Image

There is not a statistically significant effect of the display used in viewing the image when assessing viewing time ($P = 0.13$). The time taken for viewing ranged from a mean of 5.3 s for the 2-MP NEC display to a high of 5.9 s for the 3-MP GS, with the 2-MP Dell being 5.5 s and the 2-MP GS being 5.4 s.

There was a significant reader effect ($P < 0.001$), which ranged from 3.0 to 10.9 s. There was also a significant effect of day ($P < 0.001$), with a significant linear decrease in time taken as the study progressed: day 1 (6.5 s), day 2(5.5 s), day 3(5.5 s), and day 4 (4.7 s). There was also is a significant linear decrease for the order-of-reading within a single day (read number of the 1/4 of the images from 1st to 4th), $P < 0.001$, for the 1st fourth of the images viewed it took, 6.3 s, then 5.5, 5.3, and 5.1 s, respectively.

Sample Fraction Taking More than 5 s

The effect of the display used in viewing the image when assessing viewing time is of marginal significance ($P = 0.05$). The percentage of images where more than 5 s was needed to make the decision ranged from 29.5% for the 2-MP GS display to a high of 33.1% for the 3-MP GS, with the 2-MP Dell being 29.8% and the 2-MP NEC being 29.8%.

Reader also had a significant effect ($P < 0.001$), with a very wide range, from 4.6% to 75.8%. There was also a significant effect of day ($P < 0.001$), with a significant linear decrease in time taken as the study progressed: day 1(39.9%), day 2 (30.5%), day 3 (28.3%), and day 4 (23.5%). There was also is a significant linear decrease for the order-of-reading within a single day (read number of the 1/4 of the images from 1st to 4th), $P < 0.001$, for the 1st fourth of the images viewed 37.1%, 29.2%, 28.9%, and 27.0%.

Intraobserver ROC

Table 1 describes various observer results from the study. The first pair shows the sensitivity/specificity for all nine observers across the displays, as well as the average sensitivity/specificity for the specific observer and over all observers for the Display.

Table 1. Sensitivity and specificity totals by display (1, 2 = negative vs. 3, 4, 5 = positive)

Who	2 MP Dell (#1)	2 MP NEC (#2)	2 MP GS (#3)	3 MP GS (#4)	Observ. ave.
			Sensitivity		
1	74	73	78	72	74
2	88	86	88	87	87
3	97	96	97	97	97
4	77	75	80	80	78
5	92	93	96	92	93
6	64	66	68	63	65
7	86	90	91	91	89
8	86	85	88	90	87
9	92	94	95	97	94
Ave.	84	84	87	85	
			Specificity		
1	100	100	99	99	99
2	97	98	98	98	98
3	95	100	100	93	97
4	99	100	99	98	99
5	78	90	81	78	82
6	99	99	99	98	99
7	96	96	94	91	94
8	88	98	91	90	92
9	90	89	90	91	90
Ave.	94	97	95	93	

Table 2. Sensitivity and specificity *P* values by display (1, 2 = negative vs. 3, 4, 5 = positive)

Who		*P* value, 1 vs. 3	*P* value, 1 vs. 4	*P* value, 2 vs. 3	*P* value, 2 vs. 4	*P* value, 3 vs. 4
			Sensitivity			
1	0.80	0.05	0.66	<0.01	1.0	<0.01
2	1.0	1.0	0.82	0.38	0.63	0.84
3	1.0	1.0	0.73	1.0	0.55	0.75
4	1.0	1.0	0.82	0.38	0.63	0.84
5	0.68	0.05	1.0	0.21	0.65	0.08
6	0.57	0.18	0.86	0.51	0.45	0.10
7	0.14	0.03	0.01	0.61	0.45	1.0
8	0.71	0.63	0.15	0.29	0.09	0.36
9	0.15	0.08	0.01	0.77	0.30	0.58
			Specificity			
1	1.0	1.0	1.0	1.0	1.0	1.0
2	0.5	0.5	1.0	1.0	1.0	1.0
3	0.25	0.25	0.5	1.0	1.0	1.0
4	0.5	0.5	1.0	1.0	1.0	1.0
5	0.03	0.65	1.0	0.12	0.01	0.66
6	1.0	1.0	1.0	1.0	1.0	1.0
7	1.0	0.63	0.13	0.5	0.22	0.69
8	0.02	0.02	0.58	1.0	0.06	0.06
9	1.0	1.0	1.0	1.0	0.69	1.0

Table 2 compares the *P* values for sensitivity/specificity for each observer among the various Display pairings. *P* values comparing displays within a single reader are done using an exact McNemar test (sign test). Differences can be considered statistically relevant if they are below 0.05.

Of key importance to note at this juncture is the large variation in observer performance. That some of the observers performed poorly indicates that the images were not easy to interpret. Yet the fact that some of the observers perform very well indicates that the x-ray images and displays were of sufficient quality for the most expert observers to discriminate the most challenging cases. Finally, the fact that intraobserver sensitivity changed very little between the displays demonstrates that the task of detecting ILD on these images may be limited more by "anatomical noise" than display noise.

Display Measurements

Table 3 shows the minimum luminance (L_{min}), maximum luminance (L_{max}), the average ΔJND/

Table 3. Display specific measured parameters

	L_{min} (cd/m^2)	L_{max} (cd/m^2)	Average ΔJND/ΔDDL	RMS error	Ambient illuminance (lux)
Consumer color	0.58 (5%)	206 (6%)	2.06 (2%)	0.63	204 (4%)
Professional color	0.45 (2%)	213 (3%)	2.11 (2%)	0.61	193 (4%)
2 MP gray	0.77 (3%)	414 (1%)	2.42 (0.3%)	0.47	191 (3%)
3 MP gray	0.66 (2%)	402 (1%)	2.42 (0.3%)	0.31	198 (3%)

ΔDDL of the displays, the RMS error of the 256-step ΔJND/ΔDDL measurements, and the ambient light conditions of the four displays used. With the exception of the RMS error, the values in Table 1 represent the average of the periodic measurements. Corresponding errors are provided as the percent standard deviation (n, %) of the periodic measurements.

Inspection of Table 3 demonstrates that the luminance output of the Dell display varied modestly (5–6%) throughout the period of the experiment, whereas the variability of the luminance output of the remaining displays was less than 3% (including potential variation of the photometer). The grayscale displays demonstrated average ΔJND/ΔDDL values that were approximately 15% greater than that of the color displays. The RMS error relative to the display average ΔJND/ΔDDL ranged from 0.31 and 0.47 for the grayscale displays to 0.61 and 0.63 for the color displays. Repeat measurements demonstrated that the uncertainty of the RMS error measurement was less than 6%. The greater RMS error of the color displays indicates that the calibration of the color displays using OptiGrayscale was not as precise as that of the grayscale displays with SMFitAct. For further reference, note that the RMS error of all of the displays was less than the 1.0 maximum value recommended by AAPM TG 18 for primary diagnostic class displays. The ambient illuminance incident upon the displays varied by 7% or less throughout the course of the experiments.

The average coefficient of diffuse reflection of the displays was 0.0054, 0.0038, 0.0055, and 0.0038 for the consumer color, professional color, 2-MP gray, and 3-MP gray displays, respectively. Measurements acquired within the illuminance range 40–200 lux were within ±3% of the average values. Given that the room illuminance for the observer experiments was nominally 200 lux, the reflected ambient luminance was 1.08, 0.76, 1.10, and 0.76 cd/m^2 for the consumer color, professional color, 2-MP gray, and 3-MP gray displays, respectively. The diffuse reflection of ambient light elevates the effective display luminance across the entire DDL range of the display. The detrimental effects of reflected ambient light are expected to result in decreased perceived contrast in the lowest DDL, darkest regions of an image and have little affect on the highest DDL, brightest regions of an image. Given that the reflected ambient light was large compared to low DDL luminance of the displays, it is reasonable to consider whether it had an effect on the observer results. Because all of the observer results were acquired for identical ambient light conditions, this work provides no indication of whether the reflected ambient light affected observer performance. Given that the ambient light reflection properties of the displays were similar, it is not expected that they contributed to the relative observer performance on the various displays. Note that at least most of the commercially available DGSDF calibration software can include correction of for the effects of ambient light reflection. However, correction for the reflected ambient light was not included in the calibration of the displays used for this work.

Figures 1, 2, 3, and 4 show the 1D normalized noise power spectra measured using the consumer color, professional color, 2-MP gray, and 3-MP gray displays, respectively. The three spectra in each figure correspond to the nNPS measured for the CR image and display noise, display noise only, and CR image noise only. The x-axis of the power spectra were scaled to have units of (display pixel pitch)$^{-1}$. The nNPS measurements contained data up to frequencies as high as 2 cycles/pixel. The nNPS measurements acquired for frequencies greater than 1 cycles/pixel were affected by noise aliasing in the camera, which cannot be corrected for. Only the nNPS values for frequencies less than 0.5 cycles/pixel (display Nyquist) were reported here. The nNPS demonstrated many peaks corresponding to the struc-

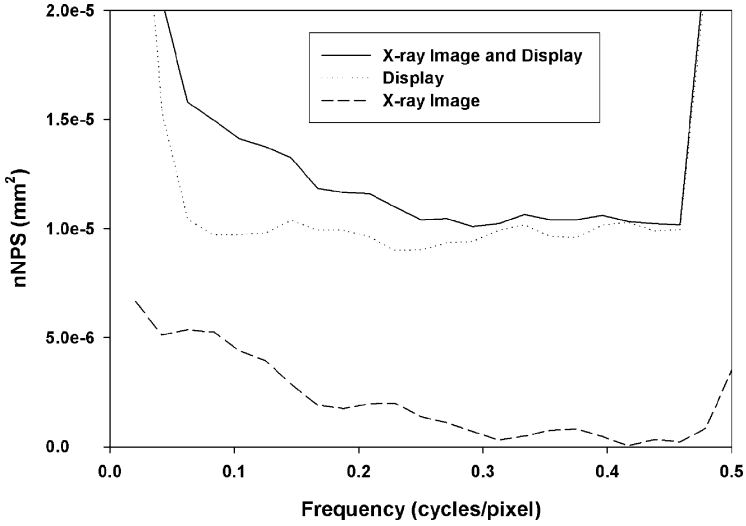

Fig 1. Normalized noise power spectra measured from the consumer color display.

tured arrangement of the image pixels and subpixels and aliasing of these frequencies by the camera sampling pitch. Most of the peaks were at spatial frequencies greater than those represented in Figures 1–4. All of the displays demonstrated an elevated low-frequency nNPS because of the relatively large area luminance nonuniformity. The magnitude of the nNPS varied by a factor of ~3.5× between the display with the least and greatest levels of display noise.

The 2D integral of the nNPS is equal to the normalized sample variance. The 2D integral of the nNPS was calculated over the frequency range −0.5 to 0.5 cycles/pixel and used as a measure of the relative noise of the various nNPS. The "variance" thus measured is equivalent to the individual display element variance. This definition of variance assumes that spatial noise with physical range that is smaller than a single pixel does not affect observer performance. The goal of the nNPS measurements was to estimate the relative contributions of display and x-ray image noise to the total noise of the displayed images. Therefore, the 2D integral of the nNPS (further referred to as the "variance") of these two contributions was normalized by dividing by the

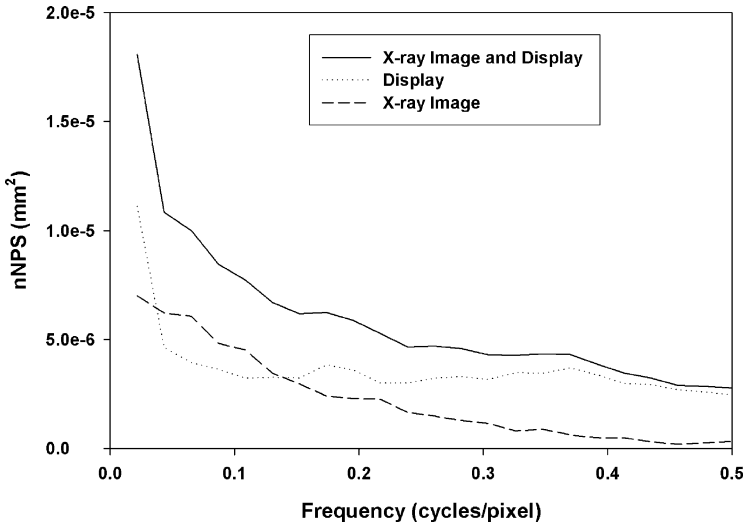

Fig 2. Normalized noise power spectra measured from the professional color display.

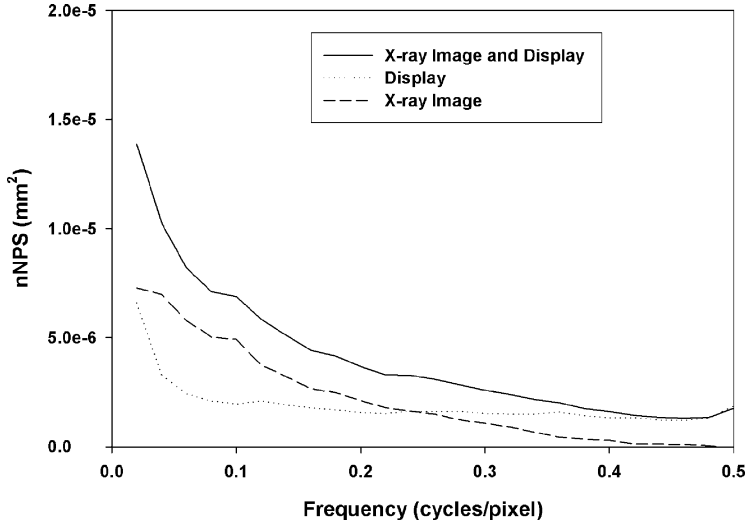

Fig 3. Normalized noise power spectra measured from the 2-MP gray display.

total variance. The relative contributions of display and x-ray image noise to the overall noise variance are provided in Table 2. The data in Table 2 indicate that all of the displays contributed a greater fraction of noise than the CR x-ray image. The relative contributions of display noise ranged from a low of 0.60 for the 2 MP gray to a high of 0.90 for the consumer color display. This wide range of display noise contributions is evident by inspection of Figures 1–4.

The nNPS results presented here should be considered preliminary in that they considered the noise of only a single x-ray image and did not attempt to account for the human visual system's perception of noise. However, these nNPS presented here demonstrate some important points about the noise of these displays. The textural noise of these displays is large with respect to the noise of a typical CR image and occupies a large absolute range. This is important because it demonstrates that the total system noise may not be limited by the noise of an x-ray imaging device, but rather by the display used to present them to an observer. For example, a decrease in

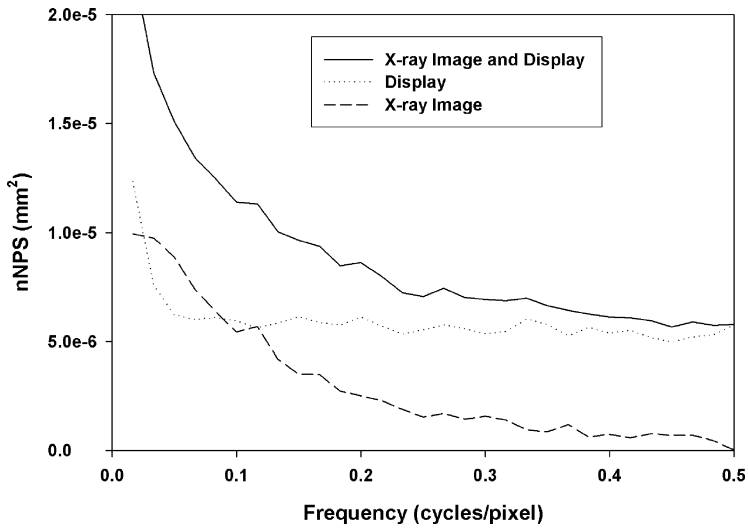

Fig 4. Normalized noise power spectra measured from the 3-MP gray display.

Table 4. Relative contribution of the display noise and the X-ray image noise to the overall variance of the displayed image

	Consumer color	Professional color	2 MP gray	3 MP gray
Display	0.90	0.73	0.60	0.78
X-ray image	0.10	0.27	0.40	0.22

x-ray image noise may not be appreciated upon display because of the relatively high display noise contribution (Tables 3 and 4).

The quality of medical images is often described by three fundamental properties: resolution, contrast, and noise. For displays, the matrix size relative to that of the images and anatomical structures to be displayed must also be considered. The following discussion will attempt to explain the (small) difference in observer experimental results in the context of the physical display performance measurements made. The displays used in this work all demonstrated good spatial resolution (low optical blur). Therefore, spatial resolution was not considered to be an important factor in this work. Display contrast (Table 1, ΔJND/ΔDDL) was approximately 15% greater for the gray displays than for the color displays. This difference might be expected to benefit the gray displays. The relative RMS error provides an indication of the ability of the manufacturer to conform to the DGSDF over a large area. That is, it is a measurement derived from the average luminance output of a large collection of adjacent pixels. The effects of the large area RMS error on the perception of the images is likely not of great importance so long as the RMS error is small relative to the ΔJND/ΔDDL "contrast", which it was for all of the displays. Of the physical measurements performed, the nNPS measurements showed the greatest difference between the displays. The magnitude of the nNPS varied by a factor of ~3.5× between the displays with the least and greatest levels of display noise. This difference in noise might be expected to influence the results of the human observer experiment. Figure 5 shows the average observer sensitivity plotted vs. the overall noise variance (display and x-ray images) for the four displays used in this study. These data indicate that the noise variance had a small influence on observer performance.

The display with the highest noise (consumer color) demonstrated the worst HVS sensitivity and the display with the lowest noise (2 MP gray) demonstrated the best HVS sensitivity. The sensitivity differences between these two displays were statistically significant. However, while the nNPS measurements varied by 350%, the observer sensitivity varied by only 3% among displays. The sensitivity of the 3-MP gray display was somewhat greater than would have been predicted

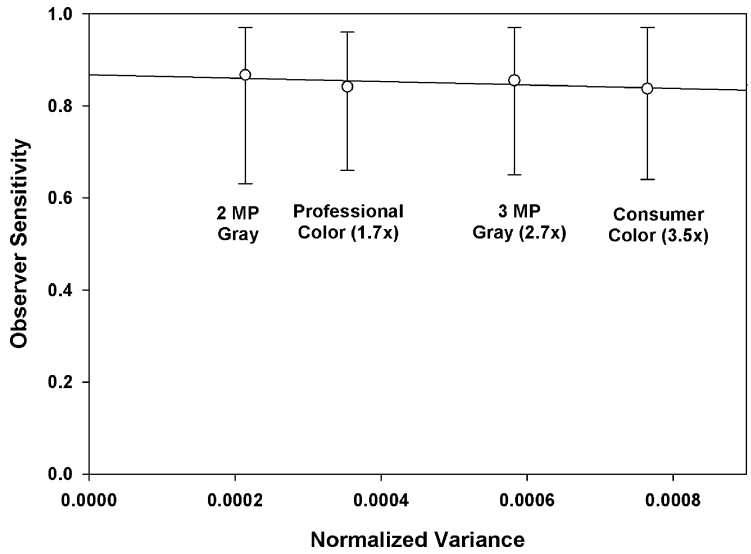

Fig 5. Average observer sensitivity plotted versus noise variance. The noise variance is the sum of the display and x-ray image noise. The value in parenthesis represents the multiplicative factor by which the variance is greater than that of the 2-MP gray display.

by the nNPS measurements. As compared to the professional color display, the 3-MP gray display may have benefited in some degree by increased contrast and a larger pixel matrix. Although these results demonstrated that minimizing display noise is beneficial to observer performance, it is not as great a factor as the large range of nNPS measurements might suggest. The discrepancy in the magnitude of change in the nNPS and human observer measurements demonstrates that there are other mechanisms in the image display and perception chain that significantly affect the ability of human observers to correctly interpret the content of the images. Likely, the most important aspect of medical imaging not included in this noise analysis is the affect of "anatomical noise" in image perception. Anatomical structures in the image, rather than x-ray image or display noise, likely represented the limiting factor in diagnosing interstitial lung disease from these images.

The results of this work seem to indicate that anatomical noise, rather than CR image or display noise, may be the most important noise source in the CR chest images used in this study. If so, one might reasonably argue that to more precisely compare display performance, images that do not contain the confounding effects of anatomical noise should be used. For example, a contrast detail (CD) study might be performed. It is reasonable to expect that a CD study may result in a greater observer performance difference between displays. However (and of course), the results of a CD (or similar) experiment cannot predict how the displays would perform for the desired task of diagnosing disease from real patient images. To accomplish that, real patient images, with the confounding effects of anatomical noise, have to be used. That this work demonstrated a minimal difference in performance between the various displays is not indicative of a limitation of the methods. Rather, it is a somewhat surprising result from a "real-world" observer experiment using images with relatively subtle evidence of disease.

CONCLUSIONS

This study concluded by finding the displays ordered as follows (in decreasing sensitivity): the 2 MP gray, 3 MP gray, 2 MP color professional class, and 2 MP color consumer class. However, it is also true that comparing the best to the worst display reveals the only statistically significant sensitivity difference.

To practitioners in the field, these results may not be surprising, but for others some explanation may be required. The perhaps antiintuitive relations among the 2- and 3-MP gray displays can be explained by noting the noise difference among those displays. As has been noted earlier, quality perception for the human visual system depends on four factors: spatial resolution, contrast resolution, blur, and noise.[8] The next point, which may be somewhat surprising, is the relatively small difference in sensitivity between the color and medical grade gray displays. This could in part be due to the conditions of this experiment, which—by matching the typical viewing conditions in an ICU or an examination room—did not permit the enhanced JND range of the gray scale displays to be totally realized. However, an additional cause could be that detection of interstitial lung disease is an anatomic noise limited, rather than display noise limited, observer task.

Physical measurements of display performance were provided and discussed in terms of their contributions or lack thereof, to differences in observer performance. Of the measurements acquired, the noise properties of the displays seemed to be indicative of observer performance. That is, increasing display noise correlated with a trend for decreasing observer sensitivity when only the 2-MP displays were considered. However, the magnitude of change in observer performance (3%) was not predicted by the magnitude of change in display and x-ray image noise (350%). Again, this likely demonstrates that the observer detection task was limited not by display and x-ray image noise, but rather by anatomical noise. Certainly, a good deal of work is required to fully understand the influences of display matrix size, resolution, contrast, and noise on observer performance.

REFERENCES

1. American Association of Physicists in Medicine Task Group 18, *Assessment of display performance for medical imaging systems* (draft), available at http://deckard.mc.duke.edu/~samei/tg18 Accessed July 8, 2003.

2. Blume H, Steven PM, Bobb M, Ho AM, Stevens F, Muller S, Roehrig H, Fan J: Characterization of high-resolution liquid–crystal displays for medical images, Medical Imaging

2002: Visualization, Image-Guided Procedures, and Display, Proc of SPIE vol. 4681 (2002).

3. Digital Imaging and Communications in Medicine, Part 14: Grayscale Standard Display Function, published by National Electrical Manufacturers Association, 2001.

4. Fetterly KA, Schueler BA: Performance evaluation of a 'dual-side read' computed radiography system. Med Phys 30:1843–1854, 2003

5. Fleiss JL, Levin B, Paik MC: Statistical Methods for Rates and Proportions, 3rd edn. New York: Wiley, 2003

6. Hsu JC: Multiple Comparisons: Theory and Methods. London: Chapman and Hall, 1996

7. Kramer CY: Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 12:307–310, 1956

8. Langer S, Bartholmai B, Fetterly K, Erickson B, Harmson S, Ryan W, Andriole K: SCAR 3'rd R&D Symposium: Efficacy of 5 megapixel CRT vs. 3 megapixel LCD display. Journal of Digital Imaging 173149–157, 2004

9. McNemar Q: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12:153–157, 1947

10. Tukey JW: The problem of multiple comparisons. In: Braun HI (Ed.) The Collected Works of John W. Tukey, Volume VIII. Chapter 1. New York: Chapman and Hall, 1994, pp 1–300