Technical Note: Gray tracking in medical color displays—A report of Task Group 196

Aldo Badano[a] and Joel Wang
Division of Imaging, Diagnostics and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland 20993

Paul Boynton
National Institute of Standards and Technology, Gaithersburg, Maryland 20899

Patrick Le Callet
Polytech Nantes/Université de Nantes, Nantes 44306, France

Wei-Chung Cheng
Division of Imaging, Diagnostics and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland 20993

Danny Deroo
Barco Kortrijk 8500, Belgium

Michael J. Flynn
Henry Ford Health System, Detroit, Michigan 48202

Takashi Matsui
Emergo, Tokyo 162-0841, Japan

John Penczek
National Institute of Standards and Technology, Boulder, Colorado 80309

Craig Revie
FFEI Limited, Hemel Hempstead, HP2 7DF, United Kingdom

Ehsan Samei
Department of Biomedical Engineering, Duke University, Raleigh, North Carolina 90281

Peter M. Steven
Independent consultant

Stan Swiderski
NEC Display Solutions, Itasca, Illinois 60143

Gert Van Hoey
Barco Kortrijk 8500, Belgium

Matsuhiro Yamaguchi
Tokyo Institute of Technology, Meguro-ku Tokyo 152-8552, Japan

Mikio Hasegawa
JVCKENWOOD Corporation, Kanagawa, 221-0022, Japan

Balázs Vince Nagy
University of São Paulo, 05508030, Sao Paulo, Brazil

(Received 15 December 2015; revised 16 May 2016; accepted for publication 19 May 2016; published 10 June 2016)

Purpose: The authors discuss measurement methods and instrumentation useful for the characterization of the gray tracking performance of medical color monitors for diagnostic applications. The authors define gray tracking as the variability in the chromaticity of the gray levels in a color monitor.
Methods: The authors present data regarding the capability of color measurement instruments with respect to their abilities to measure a target white point corresponding to the CIE Standard Illuminant D65 at different luminance values within the grayscale palette of a medical display. The authors then discuss evidence of significant differences in performance among color measurement instruments currently available for medical physicists to perform calibrations and image quality checks for the consistent representation of color in medical displays. In addition, the authors introduce two metrics for quantifying grayscale chromaticity consistency of gray tracking.

Results: The authors' findings show that there is an order of magnitude difference in the accuracy of field and reference instruments. The gray tracking metrics quantify how close the grayscale chromaticity is to the chromaticity of the full white point (equal amounts of red, green, and blue at maximum level) or to consecutive levels (equal values for red, green, and blue), with a lower value representing an improved grayscale tracking performance. An illustrative example of how to calculate and report the gray tracking performance according to the Task Group definitions is provided.
Conclusions: The authors' proposed methodology for characterizing the grayscale degradation in chromaticity for color monitors that can be used to establish standards and procedures aiding in the quality control testing of color displays and color measurement instrumentation. [http://dx.doi.org/10.1118/1.4953186]

Key words: grayscale display, GSDF, color display, grayscale tracking

1. INTRODUCTION

Color medical display systems are becoming more common in medical imaging. In some cases, color monitors are replacing grayscale monitors to accommodate more imaging modalities that the display system can handle, as well as to increase the functionality of the visualization and analysis systems. In spite of the large body of work related to grayscale displays in medical imaging, there is little understanding in the medical physics community of the methods and instrumentation available for calibrating and performing needed image quality control procedures for color displays.

A previous AAPM TG report[1] (TG18) covered some aspects of color performance of medical displays. Since the scope of TG18 was limited to grayscale displays, the color aspects included in the report amounted to the quantification of the chromaticity uniformity within the display screen and among different display devices in multiple-monitor systems. A quantitative evaluation technique was proposed using the TG18-UNL80 test pattern and a colorimeter to measure the CIE (u', v') color coordinates at the center and at the four corners of the display area of each display device. The quantity to report is the color uniformity index $\Delta(u', v')$ as the maximum distance in u', v' space between any possible pair of average u', v' points. TG18 establishes that, based on clinical experience, a color uniformity parameter of 0.01 or less is necessary to assure acceptable color matching of primary class grayscale display devices of a workstation. In addition, the report recommends that the distance between any pair of color coordinates across the display area of each device should also not exceed the 0.01 limit. These criteria apply to primary class displays while no quantitative requirements are specified for secondary class displays.[1]

More recently, a number of reports have highlighted the relevance of performing colorimetric measurements on medical display.[2–6] However, no work has been reported on practical methodologies for the assessment of the color presentation states of a medical display in the clinical setting. To address this need, the AAPM formed TG196 with the aim of developing professional documents for guiding the medical physicist in his or her clinical duties. In this report, we discuss methods for the characterization of the gray tracking performance of medical color monitors for diagnostic applications. We first discuss measurement methods and instrumentation for gray tracking performance of color monitors and present and discuss data regarding the capability of color measurement instruments with respect to their abilities to measure a target white point corresponding to the CIE Standard Illuminant D65 at different luminance values within the grayscale palette of a medical display. The analysis shows significant differences in performance among color measurement instruments currently available for medical physicists for performing calibrations and image quality checks for the consistent representation of color in medical displays. Finally, we introduce and propose methods and tolerance limits for gray tracking of medical display systems based on achievable levels of performance and conclude about the impact of these findings on the requirements for quality control of color medical monitors with respect to the changes in chromaticity across the grayscale.

2. INSTRUMENTATION

For the purpose of this work, we define two classes of color measurement instrumentation. First, we define the reference instrument, typically a spectroradiometer with a wide sensitivity range and high accuracy and precision, calibrated or traceable to a secondary or primary standard laboratory. Secondly, we define a field instrument which is typically a colorimeter system. Some of which use filters to measure the color of the light emitted by the display. The typical cost of a field instrument ranges from $100 to $1500 while a reference instrument costs over $10 000.

3. INTERCOMPARISON

In order to determine the capabilities of available instrumentation and determine reasonable tolerances to set for the color characteristics of a medical monitor in the context of the clinical tasks of a medical physicist, we performed an intercomparison experiment involving several medical display laboratories. For this experiment, a display device BARCO Nio Color 3MP (MDNC-3121) with a CCFL backlight and computer were sent to the participating laboratories over the course of 18 months.

For the intercomparison study, a wide array of color measurement instruments were used to collect color data. The instruments ranged from high-end spectroradiometers to consumer-grade color calibration kit (CCK) meters. Instruments came from various manufacturers, including Konica Minolta, TOPCON, Instrument Systems, X-Rite, and Datacolor. We classified the Konica Minolta CS2000 spectroradiometer, TOPCON SR-3 spectroradiometer, Instrument Systems CAS-140D (TOP200-100 Optical Probe) spectrometer, and Konica Minolta CS1000 spectroradiometer under our reference meter classification. We classified the Konica Minolta CS100A luminance and color meter, Konica Minolta CA310 color analyzer, X-Rite i1 Pro CCK meter, X-Rite Chroma 5 colorimeter, and Datacolor Spyder 4 CCK meter under our field classification.

3.A. Protocol

Participating laboratories were instructed to measure the display device with their reference meter/sensor, as well as, lower cost colorimeters used in the field using the following protocol. For this study, only one meter per laboratory was categorized as a reference meter so as to not skew the results of the study toward a particular lab. There were a total of 6 laboratories who participated in the intercomparison study.

Participants included one lab from Belgium, three labs from Japan, one lab from Brazil, and one lab from the United States. Each participating laboratory was asked to unpack the PC and the display and leave the equipment in the measurement room for a minimum of 1 h to adapt to the ambient temperature. The measurement room was defined to be a dark room with an illuminance of less than 5 lux at the faceplate of the monitor. After the 1-h wait, the PC was to be connected to the display and turned on to verify that all the OSD settings are correct according to the protocol specifications. Then, the display was to warm up for a minimum of 1 h while recording the room temperature.

The computer was preloaded with images of 48 color patches representative of grayscale and color channels. The grayscale patches are at intervals of 15 levels starting from 0 to 255. The red, green, and blue values were a smaller subset with intervals of 15 and 30 (15, 30, 45, 75, 105, 135, 165, 195, 225, and 255). An image viewer (bitmap viewer) on the PC was opened to view the patterns. The color measurement meter was then aligned to measure the color pattern image. In order to ensure display stability, a repeatability challenge was posed on certain gray level patterns by measuring each required pattern 5 times and documenting that the variability was within acceptable limits. To qualify the experiment, several conditions had to be met. The variation in Y color coordinate (CIE 1931, 2° observer) had to be less than 10% for level 0, less than 2% for levels 120 and 255, the (u', v') variation had to be less or equal than 0.005 for level 0, and the (u', v') variation had to be less or equal than 0.001 for levels 120 and 255. When the repeatability conditions were met, manual measurements were taken and recorded for each of the 48 color patches.

For all measurements, participants were asked to iterate through several steps to capture the measurement data. The test pattern was displayed on the monitor. Then, after 5 s, they were asked to perform the measurement and record the CIE tristimulus XYZ values in the measurement worksheet, and then to repeat this procedure for the rest of the test patterns. After completing the data collection, each lab sent its results to the FDA for aggregation, anonymization, and analysis.

3.B. Data analysis

In order to represent the data for visualization, we report various plots to help compare various instruments across different countries and laboratories. For each color pattern, we defined a reference point (u'_R, v'_R) which was calculated by finding the average X, Y, and Z components of designated reference instrument measurements, and then converted into (u', v') coordinates.

For each color pattern, a 1D representation of Euclidean (u', v') distance, also known as the chromaticity difference, between each measurement and the defined reference point

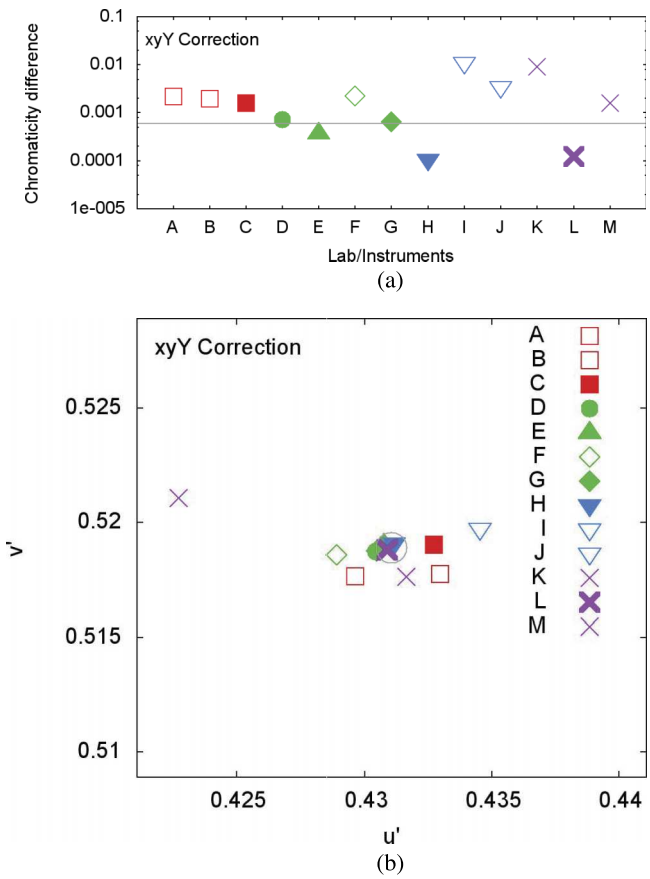

FIG. 1. Results of the intercomparison study. (a) Variation in the Red105 color pattern measurements. Each point represents the distance in the (u', v') space from a measurement from an instrument/lab to the reference point, the colors represent different countries, and the solid (filled-in or bolded) points are classified as reference meters. (b) Same data in a 2D representation for the Red105 color pattern measurements. Each point represents a measurement from an instrument in the (u', v') space, centered at the reference point. The gray line/circle represents the average distance between the reference point and reference meter measurements.

chosen in this study was generated. Figure 1(a) shows the 1D representation for the red level 105 pattern. Using these charts, we can quantify how each instrument performed compared to others across different countries and instrument classes. In addition, 2D representations were generated where we plot measurement results on the (u', v') plane centered at the reference point, with a gray circle with a radius corresponding to the average chromaticity difference between the reference meter measurements and the reference point, analogous to the gray line in the 1D plots. Figure 1(b) shows the 2D representation for the red level 105 pattern. During the analysis, we noticed a systematic temporal drift in color of the monitor used as a reference over the course of the 14 months that took to complete the entire intercomparison study. To compensate for this drift, we fitted the drifted results using a model, corrected the data in the (x, y) color space, and converted it all back to a (u', v') space.[7] All data presented in the main section of this report have been corrected for the display's temporal drift.

3.C. Results

Average chromaticity difference charts were generated to compare how instruments performed with respect to color and instrument class (reference versus field). The general trend suggests that the reference meters perform at a different level than the field instruments (full intercomparison data sets are available in Ref. 7). We observe in Fig. 2, a general trend where the average chromaticity difference decreases as the digital count, which is correlated with luminance, increases. In addition, we note that the average chromaticity difference for field instruments is much larger than that of reference instruments for all levels.

In summary, the results suggest that the average distance observed for reference meters is equal or less than approximately 0.001, while for field instruments, it is at an order of magnitude higher, or approximately 0.01. As the field instrument class contains devices with a wide range of specifications, higher performance devices within the class can achieve uncertainties smaller than 0.01.

4. GRAY TRACKING METHODS

In this section, we propose two different methods that can be used, in conjunction with appropriate instrumentation, to determine the variation in color of a grayscale gradation. We assume here a calibrated display to the grayscale standard display function (GSDF). We also assume for the examples presented in this work a target white point color coordinates corresponding to D65, although it has to be noted that the recommendation of a specific white point target (e.g., D65) is not yet established in the community and is beyond the scope of this report. Future work will address the setting of the white point in connection to the rest of the settings that would define the performance of the display color presentation.

Using TG18-LN test patterns (TG18-LNx, $x = 01,02,\ldots,18$) and with a color meter, we record the luminance and color coordinates (u',v'). For the analysis, we only consider measurements from the set of 18 patterns with a luminance value higher than the maximum of either 5 cd/m^2 or 1% of the maximum luminance measured for TG18-LN18 (full white) (Fig. 3). We denote the number of remaining measurements as N.

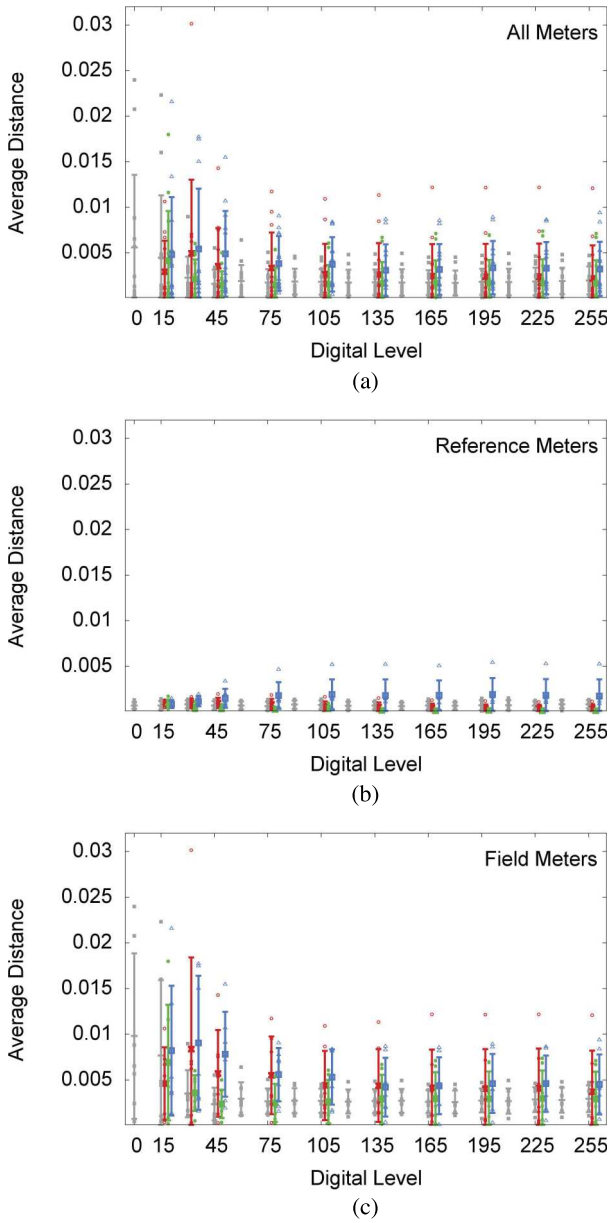

FIG. 2. Average chromaticity difference between a set of measurements and the reference point across (a) all measurements produced by all meters, (b) measurements produced by only reference meters, and (c) measurements produced by only field meters. Chromaticity differences between the reference point and measured pattern for each instrument are represented as points. The error bars (standard deviation of chromaticity differences) are centered at the average chromaticity difference of measurements and the reference point made by the above defined set of meters for each pattern. The colors represent their corresponding color patterns.

We define the quantity of interest ($\mathcal{T}_1$) as the average distance in the (u', v') plane between the measurement of full white (TG18-LN18) and the nondiscarded measurements

$$\mathcal{T}_1 = \frac{1}{N-1} \sum_{j=17-N+1}^{17} \Delta_1(u', v'), \tag{1}$$

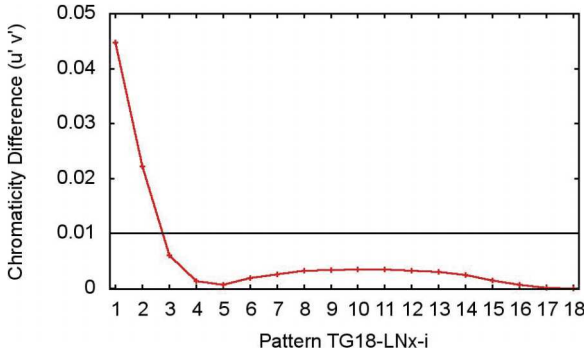

FIG. 3. Chromaticity difference between each of the patterns and the full white (TG18-LN18).

where

$$\Delta_1(u', v') = \sqrt{(u'_j - u'_{18})^2 + (v'_j - v'_{18})^2}. \quad (2)$$

This metric quantifies how close the grayscale chromaticity is on average to the chromaticity of full white, with a lower value representing an improved grayscale tracking performance. Table I illustrates the calculation of the quantity for the display used in the intercomparison, based on measurements with one of the reference sensors from the intercomparison study. The value of $\mathcal{T}_1$ in this example is 0.0023.

Alternatively, one can define the grayscale tracking metric as the average distance in the (u', v') plane between consecutive (nondiscarded) measurements ($\mathcal{T}_2$),

$$\mathcal{T}_2 = \frac{1}{N-1} \sum_{j=17-N+1}^{17} \Delta_2(u', v'), \quad (3)$$

TABLE I. Measurement data of the TG18-LNx patterns for the calculation of $\Delta(u'v')$. The corresponding values for the metrics $\mathcal{T}_1$ and $\mathcal{T}_2$ are 0.0023 and 0.0005.

Pattern	L (cd/m^2)	L (%)	u'	v'	$\Delta_1(u', v')$	$\Delta_2(u', v')$
TG18-LN01	0.42	0.10	0.1948	0.4281	0.0440	N/A
TG18-LN02	0.82	0.20	0.1998	0.4495	0.0220	0.0219
TG18-LN03	2.26	0.60	0.2039	0.4649	0.0060	0.0160
TG18-LN04	4.25	1.10	0.2047	0.4694	0.0014	0.0046
TG18-LN05	6.69	1.70	0.2048	0.4715	0.0006	0.0020
TG18-LN06	10.2	2.50	0.2049	0.4727	0.0019	0.0012
TG18-LN07	14.7	3.70	0.2050	0.4735	0.0026	0.0008
TG18-LN08	21.0	5.20	0.2051	0.4740	0.0032	0.0006
TG18-LN09	29.5	7.40	0.2051	0.4743	0.0034	0.0002
TG18-LN10	41.3	10.3	0.2051	0.4744	0.0035	0.0001
TG18-LN11	57.6	14.3	0.2053	0.4743	0.0035	0.0002
TG18-LN12	77.7	19.3	0.2051	0.4741	0.0032	0.0003
TG18-LN13	101.3	25.2	0.2052	0.4738	0.0030	0.0003
TG18-LN14	131.1	32.7	0.2053	0.4733	0.0025	0.0005
TG18-LN15	177.7	44.3	0.2050	0.4724	0.0015	0.0010
TG18-LN16	235.1	58.6	0.2049	0.4715	0.0007	0.0009
TG18-LN17	304.7	75.9	0.2049	0.4708	0.0001	0.0007
TG18-LN18	401.5	100	0.2050	0.4708	0.0000	0.0001

where

$$\Delta_2(u', v') = \sqrt{(u'_j - u'_{j+1})^2 + (v'_j - v'_{j+1})^2}. \quad (4)$$

This metric quantifies how well the grayscale chromaticity is kept constant throughout the grayscale range, with a lower value representing also an improved grayscale tracking performance. The value of $\mathcal{T}_2$ in this example is 0.0005.

It has to be noted that $\mathcal{T}_1$ and $\mathcal{T}_2$ represent average deviations from the desired behavior of the grayscale. It is useful in practice to define the relevant metric in terms of the maximum deviation depending on the desired way to capture deviations either with respect to the white point chromaticity or between consecutive measurements. We define $\mathcal{T}_{\max}$ as the maximum deviation in $\Delta(u'v')$ for the nondiscarded measurements, as follows:

$$\mathcal{T}_{i,\max} = \max(\Delta_i(u'v')_j), \quad (5)$$

where j is the number of nondiscarded measurements included in the analysis. In our example case, the value of $\mathcal{T}_{\max}$ is 0.0035 if we consider deviations from the white point chromaticity ($\mathcal{T}_{1,\max}$) and 0.0013 if we consider deviations between consecutive measurements ($\mathcal{T}_{2,\max}$).

The usage of these metrics needs to be determined considering the significance for clinical practice. $\mathcal{T}_{1,\max}$ emphasizes a constant deviation of the chromaticity throughout the grayscale range with respect to the chromaticity of full white. $\mathcal{T}_{2,\max}$ relates to the presence or absence of abrupt chromaticity changes between consecutive grayscales which might be more visually significant than gradual changes of chromaticity. Generally, if the value of $\mathcal{T}_{1,\max}$ is small, $\mathcal{T}_{2,\max}$ will also be small and might not provide significant additional information. If $\mathcal{T}_{1,\max}$ is large, then the value of $\mathcal{T}_{2,\max}$ provides an indication of whether the behavior implies gradual or abrupt chromaticity changes.

Defining acceptable values for the gray tracking metrics is beyond the scope of this effort and requires additional work to understand the effect of these changes on diagnostic performance. Nevertheless, it is of interest to note that the TG18 report[1] puts forward an upper limit of 0.01 for a spatial chromaticity evaluation on test pattern TG18-UNL80. Beyond the perceptual effects of gray tracking, the proposed metrics are useful for device characterization when performing acceptance testing and for quality control procedures.

5. CONCLUSION

As more medical imaging systems begin to incorporate color monitors, it becomes essential to develop methods of characterizing the grayscale tracking performance of displays using appropriate color measurement instrumentation. Our findings suggest the capabilities and limitations of various color measurement instruments. Our proposed methods for characterizing the grayscale degradation in chromaticity for color monitors can be used to establish standards and procedures aiding in the

quality testing of color displays and color measurement instrumentation.

ACKNOWLEDGMENTS

Special thanks to the colleagues of Colorimetry Laboratory of CPqD, Brazil (Celso Pinto Saraiva and Antonio Umberto Pedrazzani, Junior) and to the Photometry Laboratory of IEE-USP (Jos Gil Oliveira, Marcio Ribeiro, Marcelo Jesus, and Rinaldo C. Pinto). The mention of commercial products herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the Food and Drug Administration and is not subject to copyright.

CONFLICT OF INTEREST DISCLOSURE

The authors have no COI to report.

[a]Author to whom correspondence should be addressed. Electronic mail: aldo.badano@fda.hhs.gov.

[1]E. Samei *et al.*, "Assessment of display performance for medical imaging systems: Executive summary of AAPM TG18 report," Med. Phys. **32**, 1205–1225 (2005).

[2]J. Fan, H. Roehrig, and E. Krupinski, "Characterization of color-related properties of displays for medical applications," in *Medical Imaging, International Society for Optics and Photonics* (SPIE, 2007), p. 65160X.

[3]E. A. Krupinski, H. Roehrig, and T. Matsui, "High luminance monochrome vs color displays: Impact on performance and search," Proc. SPIE **7966**, 79661R-1–79661R-6 (2011).

[4]H. Roehrig, K. Rehm, L. D. Silverstein, W. J. Dallas, J. Fan, and E. A. Krupinski, "Color calibration and color-managed medical displays: Does the calibration method matter?," Proc. SPIE **7627**, 76270K-1–76270K-13 (2010).

[5]L. D. Silverstein, S. F. Hashmi, K. Lang, and E. A. Krupinski, "Paradigm for achieving color-reproduction accuracy in LCDs for medical imaging," J. Soc. Inf. Disp. **20**(1), 53–62 (2012).

[6]W.-C. Cheng and A. Badano, "Virtual display: A platform for evaluating display color calibration kits," SID Symp. Dig. Tech. Pap. **42**, 1030–1033 (2011).

[7]See supplementary material at http://dx.doi.org/10.1118/1.4953186 for details.