Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Does preference influence performance when reading different sizes of cranial computed tomography?

Antje C. Venjakob
Tim Marnitz
Lavier Gomes
Claudia R. Mello-Thoms

Does preference influence performance when reading different sizes of cranial computed tomography?

Antje C. Venjakob,^{a,†,*} Tim Marnitz,^{b,†} Lavier Gomes,^c and Claudia R. Mello-Thoms^{d,e}

^aTechnische Universität Berlin, Department of Psychology and Ergonomics, Chair of Human-Machine Systems, Marchstrasse 23, Berlin 10587, Germany

^bCharité Universitätsmedizin Berlin—Campus Virchow Klinikum, Klinik für Strahlenheilkunde, Mittelallee 3, Berlin 13353, Germany

^cWestmead Hospital, Corner of Darcy Road & Bridge Street, Westmead, Sydney, NSW 2145, Australia

^dUniversity of Sydney, Medical Imaging and Radiation Sciences, Faculty of Health Science, 94 Mallet Street, Level 2, Room 204, Sydney, NSW 2150, Australia

^eUniversity of Pittsburgh, Department of Biomedical Informatics, 5607 Baum Boulevard, Room 423, Pittsburgh, Pennsylvania 15206-3701, United States

Abstract. Radiology practice is based on the implicit assumption that the preference for a particular presentation mode goes hand in hand with superior performance. The present experiment tests this assumption in what pertains to image size. Forty-three radiologists were asked to identify intracranial hemorrhages on 20 cranial computed tomography scans in two image sizes, 14 × 14 and 28 × 28 cm. They were asked to indicate which size they preferred and subsequently rated each size on a continuous scale in terms of how much they liked them. The results show no correlation between the jackknife free-response receiver operating characteristic figure of merit and preference rated on a continuous scale (large image: $r = 0.14$, $p = 0.38$; small images: $r = 0.14$, $p = 0.39$). Similarly, there was no significant correlation between the time a radiologist took to read a case and preference rated on the continuous scale (large image: $r = -0.07$, $p = 0.64$; small images: $r = -0.04$, $p = 0.80$). When dividing radiologists into two groups according to their size preference, there was no significant difference in performance between groups with regard to either large or small images. The results suggest that the preference for an image size and performance with regard to it are not related. © 2014 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.1.3.035503]

Keywords: image size; preference; observer performance; cranial computed tomography.

Paper 14124PR received Sep. 15, 2014; accepted for publication Nov. 17, 2014; published online Dec. 22, 2014.

1 Introduction

In the modern digital radiology clinic, numerous settings can be adjusted before a case is interpreted. These settings comprise not only the hardware of the workstation, such as the luminance of the monitor, but also the digital presentation of the images. For example, radiologists can set the contrast to a desired level, and the image size can be adjusted before as well as during the process of interpretation.

The number of settings that can be adjusted individually generates a combinatorial explosion, which prevents the formulation of guidelines regarding optimal parameterization to foster improved diagnostic performance. To date, there is only a very limited number of studies that have focused on the effects of image size and its importance for diagnostic performance.^{1–5} Interestingly, the results of these studies are mixed.

One of the first image size experiments in digital computed tomography (CT) focused on the comparison of four differently sized chest CTs containing nodules, lines, and micronodular opacities.¹ The size of the images ranged from 15 × 19 to 35 × 43 cm. It was found that the interpretation accuracy varied depending on the type of lesion; while detection of nodules was not affected by image size, detection of lines and micronodular opacities declined with decreasing image size. However, in an experiment that compared the effect of image size with the

detection of lung nodules in chest CT², a significant reduction in reading accuracy was observed for smaller images when participants were not able to compensate for size by moving closer to the image. Although in both studies the CT scans were acquired digitally, they were presented to the interpreters as hardcopies in a tiled display format.

Gur et al.³ found a slight advantage for performance, measured by the area under the receiver operating characteristic (ROC) curve, as well as for reading time for small (12 × 12 cm) compared with large (24 × 24 cm) abdominal computed tomographies shown in stack mode, where radiologists' searched for lesions in the patients' livers. Interestingly, the participants nonetheless rated the large images as more comfortable to interpret. Yamaguchi et al.⁴ also found a statistically significant advantage for small images (13 × 13 cm) compared with enlarged ones (30 × 30 cm) regarding the area under the ROC curve when identifying nodular ground-glass opacity on CT displayed in cine mode. Although Yamaguchi et al. did not report a systematic evaluation of comfort, they mentioned that all seven interpreters strongly indicated that the original image size of 13 × 13 cm was easier to interpret. No differences in reading time were found.

The seemingly contradictory results of these studies are puzzling. One hypothesis may be that the differing results are due to the differences between the study designs with regard to other factors that are involved in image perception such as

*Address all correspondence to: Antje C. Venjakob, E-mail: antje.venjakob@mms.tu-berlin.de

[†]These authors share first-authorship of this paper.

hard- versus softcopy readings, stack- versus tile-mode, the specific image sizes that were used, or the type of lesion presented to the radiologist. However, another possibility may be that no unambiguous effects of image size are found because performance is not so much determined by the external image characteristics, but rather it is a function of the individual reader.

In summary, many studies in medical image perception have implicitly laid a link between radiologists' preference and performance, with some studies using preference as the sole indicator of superiority of one modality.⁶ However, it is not clear if there is indeed a connection between the two. With digital imaging, the possibilities to adjust image size and other parameters have increased enormously, and size can be set according to individual preference easily. Hence, if individual preference is a valid indicator of performance, then this would be a valuable finding. On the other hand, it would be even more important to know if preference was not directly related to performance. In this case, scientifically established guidelines could replace individual preference in the reading room. The aim of this paper is to explore if there is a correlation between radiologists' preferred image size and their performance. To this end, performance and preference for two image sizes will be assessed separately. Diagnostic accuracy, as measured by the jackknife free-response receiver operating characteristic (JAFROC) figure of merit (FOM), and reading time for each of the two image sizes will serve as indicators of performance. Performance and reading time will be correlated to a continuously rated preference score. To gain insight into the reasons underlying preference for a particular image size, the motivation for a preference will be assessed, too.

2 Materials and Methods

2.1 Participants

The data of a total of 43 radiologists contributed to the analysis reported in this paper. The data were collected in two different University hospitals, one in Germany and one in Australia. In Germany, 21 radiologists participated, and in Australia, 22 radiologists were recruited. Demographic characteristics of the two samples are summarized in Table 1. The data show that the participants in Australia were more experienced than the ones in Germany.

2.2 Cases

For the study, 10 cranial CT (cCT) cases, with a slice thickness of 5 mm and a total of 25 to 32 slices per case, were selected. These cases were rated as being normal independently by three radiologists. They were deidentified, i.e., all patient information

was removed and converted to noncompressed .png images. The images comprised 512×512 pixels, which amount to 14×14 cm. This is the original image size, and it was included in the study as "small images," as any size smaller would be accompanied by a loss in diagnostic information. The cases were flipped around their vertical axis, resulting in 20 seemingly different cases. A total of 18 subtle hemorrhages were inserted into 10 of the cases using the open-source graphics program GIMP. The lesions spanned between one and three slices and were taken from other, more severe cases that featured numerous hemorrhages. Subsequent to this, the images of all 20 cases were enlarged to a size of 1024×1024 pixels (28×28 cm) using the cubic Catmull-Rom splines interpolation algorithm.⁷ This is the largest size that enables the complete display of the image on the monitor used. The images of each case were then inserted into otherwise black slides of a Microsoft Powerpoint presentation to enable the radiologists to scroll back and forth through the stack using a mouse wheel. An example of a stack-mode cCT case with an inserted hemorrhage and the reporting of one radiologist is displayed in Fig. 1.

2.3 Experimental Design

The displayed image size served as an independent variable in this experiment. Two different image sizes were presented to the radiologists: "large images" (28×28 cm) and "small images" (14×14 cm). Image size was implemented as a within-subjects factor in the study, i.e., each radiologist reads 10 small and 10 large cases. Each single case was presented either as small or as large image to participants; hence, no participant saw the same case twice. A binary preference rating that the participants provided served as a second independent variable. Preference constitutes a between-subjects factor with two levels, which are "preference for large images" and "preference for small images."

The dependent variables of the experiment were performance, reading time, and preference. Performance was measured as the JAFROC FOM using a scale from 1 (very little confidence in decision) to 10 (very confident that the encircled structure indeed represented an intracranial hemorrhage). The time that the radiologists took to complete the reading of a single case was assessed by recording the interval between the display of the first slice and the closure of the case. Using an image size questionnaire at the end of the experiment, radiologists were asked which of the two image sizes they preferred. This was done by posing the question, "Which image size do you prefer?" Next to the question, readers could either tick a box for large or for small images and answer an open question indicating why they preferred the indicated size. Radiologists were additionally asked to rate their liking for each image size on a continuous scale from 0 (dislike) to 10 (like). They were also asked if they would have preferred a different size and if they perceived image size to be an important factor in their diagnostic interpretation of the case.

2.4 Procedure

All radiologists participated during their working hours. They read and signed an informed consent form and read the specific instructions to the experiment on the computer's screen. They were instructed to only search for intracranial hemorrhages. They were asked to encircle a potential hemorrhage using the mouse as soon as they decided that they would report it. The radiologists were also asked to indicate their confidence in their decision by writing their confidence next to their marking.

Table 1 Demographic characteristics of the participating radiologists of the two facilities. Mean values are shown (standard deviations are displayed in parenthesis).

	Germany	Australia
Number of radiologists	21	22
Age	34.3 years (6.5)	40.5 years (11.8)
Years of experience	6.4 years (6.3)	12.5 years (13.7)
Female/male	6/15	4/18

Fig. 1 Cranial computed tomography (cCT) case in stack-mode imaging. On the top slice, an inserted hemorrhage is visible along with the marking made by one of the participating radiologists. Radiologist's confidence rating of "9" is also shown.

The reading started with a practice CT during which the radiologists were given the opportunity to ask questions. The subsequent data collection followed in two blocks: half of the participants were shown 10 small cases first, whereas the other half started with 10 large cases. After a break of maximally 5 min, the participants read the cases of opposite size. There was no time limit for the interpretation of the cases. Upon completing the reading of the 20 cases, all participants filled out the image size questionnaire.

2.5 Data Analysis

We hypothesized that the data of the two institutions would be similar. To ascertain whether this is the case, a 2×2 analysis of variance was performed with image size and institution serving as independent variables and continuously rated preference, JAFROC FOM and reading time being the dependent variables. It was decided to analyze the data conjointly for the two institutions if no interaction between image size and institution occurred. Further, two-way analyses of variance were used to compare the influence of the two independent variables, displayed image size and preference, on continuous preference, JAFROC FOM, and reading time. An interaction between the two factors regarding the continuous preference rating is assumed to signal internal consistency between the two preference measures, while a significant interaction effect regarding the JAFROC score or reading time would suggest a relationship between the preference and performance measures. In addition, bivariate correlations between preference and performance were calculated using Pearson correlations. With regard to reasons for the preference of an image size, only descriptive statistics will be reported. Inferential statistics were calculated using JAFROC software⁸ and IBM SPSS version 22. All tests are two-tailed, and the alpha level of significance was set to $p < 0.05$.

3 Results

3.1 Comparability of Institutions

The comparison of the two institutions yielded a significant main effect with reading time being significantly longer in the Australian sample compared with the German. However, no significant interaction was shown for any of the three dependent variables [continuously rated preference: main effect image size, $F(1,41) = 1.47$, $p = 0.23$, $\eta^2 = 0.04$, main effect institution, $F(1,41) = 0.18$, $p = 0.67$, $\eta^2 = 0.004$, interaction effect, $F(1,41) = 0.03$, $p = 0.86$, $\eta^2 = 0.001$; JAFROC FOM: main effect image size, $F(1,41) = 0.35$, $p = 0.56$, $\eta^2 = 0.009$, main effect institution, $F(1,41) = 0.08$, $p = 0.78$, $\eta^2 = 0.002$, interaction effect, $F(1,41) = 0.41$, $p = 0.53$, $\eta^2 = 0.01$; reading time: main effect image size, $F(1,41) = 3.75$, $p = 0.06$, $\eta^2 = 0.08$, main effect institution, $F(1,41) = 8.5$, $p = 0.01$, $\eta^2 = 0.17$, interaction effect, $F(1,41) = 0.39$, $p = 0.53$, $\eta^2 = 0.009$]. The data were therefore analyzed conjointly for the two institutions.

3.2 Preference

When asked the binary question of what image size they preferred, 22 radiologists indicated that they preferred the large images, and 16 radiologists showed a preference for small images (Australia: preference for large images: 12, small images: 8, no preference: 2; Germany: preference for large images: 10, small images: 8, no preference: 3). Five radiologists did not use the predefined answers and wrote next to them that they did not have a preference. According to their binary preference rating, radiologists were divided into two groups: those who prefer small and those who prefer large images. The mean values and standard deviations for the two groups and the dependent variables are displayed in Table 2.

The 2×2 analysis of variance showed no significant main effect of the displayed image size, $F(1,36) = 0.23$, $p = 0.64$, $\eta^2 = 0.006$, or of the preferred image size regarding the continuous preference rating, $F(1,36) = 1.29$, $p = 0.26$, $\eta^2 = 0.04$. There was, however, a significant interaction of preferred and

Table 2 Mean continuous preference rating, JAFROC figure of merit (FOM), and reading time split for the two preference groups regarding their binary preference score (standard deviations in parenthesis).

Binary preference	Continuous preference rating, large images	Continuous preference rating, small images	JAFROC FOM, large images	JAFROC FOM, small images	Reading time in seconds, large images	Reading time in seconds, small images
Large images	8.0 (1.3)	3.7 (1.9)	0.69 (0.05)	0.70 (0.05)	68.5 (26.6)	65.9 (17.4)
Small images	4.0 (2.4)	7.3 (2.3)	0.68 (0.06)	0.67 (0.05)	75.1 (30.1)	63.0 (26.0)

displayed image sizes, $F(1,36) = 77.6$, $p < 0.001$, $\eta^2 = 0.68$, because radiologists who preferred large images in the binary rating rated these better in the continuous rating. Similarly, radiologists who preferred small images in the binary rating preferred these in the continuous rating over large images.

3.3 Performance

The 2×2 analysis of variance regarding the JAFROC scores yielded no significant main effect for the displayed image size, $F(1,36) = 1.00$, $p = 0.33$, $\eta^2 = 0.03$, or the preferred image size, $F(1,36) = 0.50$, $p = 0.83$, $\eta^2 = 0.001$, nor did it yield a significant interaction between the two, $F(1,36) = 0.49$, $p = 0.49$, $\eta^2 = 0.01$. Regarding reading time, there was a significant main effect for the displayed image size with small images being read significantly faster than large ones, $F(1,36) = 4.68$, $p = 0.04$, $\eta^2 = 0.12$. There was no main effect for the preferred image size, $F(1,36) = 0.06$, $p = 0.80$, $\eta^2 = 0.002$, and the interaction between the two factors did not reach significance either, $F(1,36) = 1.94$, $p = 0.17$, $\eta^2 = 0.05$.

3.4 Relationship Between Preference and Performance

Table 2 shows the descriptive statistics for continuous preference rating, JAFROC FOM, and reading time, divided according to preference for small or large images.

No significant bivariate correlations could be found between the continuous preference rating for an image size and performance with regard to it. This was true for performance measured in terms of diagnostic accuracy, i.e., the JAFROC FOM (large images: $r = 0.14$, $N = 43$, $p = 0.38$; small images: $r = 0.14$, $N = 43$, $p = 0.39$), as well as for reading time (large images: $r = -0.07$, $N = 43$, $p = 0.64$; small images: $r = -0.04$, $N = 43$, $p = 0.80$).

3.5 Radiologists' Reasons for Preference Choices

The radiologists' motivation for their choice of preferred image size and the importance they saw in the topic are shown in Table 3.

4 Discussion

In this study, we examined whether preference and performance are related when reading cranial stack-mode CT. We grouped participants by their reported preference for large and small images and analyzed their continuous preference rating and performance, as measured using the JAFROC FOM and reading time. No interaction was found regarding radiologists' preferences for an image size and detection of hemorrhages when using that particular image size. Similarly, when correlating

radiologists' continuous preference scores to performance and reading time, no significant correlations were found for either small or large images. Together, the results of this study suggest that there is no relationship between the preference for an image size and the radiologists' performance when reading cases in their preferred size.

Although results that show no significant difference need to be interpreted with care, they, nonetheless, suggest that the preference may not be a sufficient criterion when it comes to the selection of image size, and there is no reason to assume that these findings are exclusive to image size. Though not tested in this study, it is likely that in the process of image interpretation, the selection of other software and hardware settings should not be based on preference ratings alone, as these cannot be assumed to imply superiority in performance. As no significant differences in performance could be found when comparing readings on 14×14 to 28×28 cm images, it can be assumed that the current practice of radiologists choosing their preferred image size is not harmful. However, no benefits can be expected from the current process either.

The missing link between preference and performance does not mean that there is no room for improving reading performance, nor does it mean that the assessment of preference is useless. However, the way to improvement is clearly more laborious than guidance by individual preference. Different image sizes could be advantageous for different phases in the interpretation process. This is suggested by the statements of five radiologists, who said that they liked small images for a good overview early on in the interpretation process, which helped them detect possible hemorrhages, and liked using larger images to take a detailed look at structures they had identified as potential perturbations. These findings highlight the need for a more holistic approach to the study of image size. Such an approach should include measures that enable the study of perceptual and cognitive processes, such as eye tracking, along with performance and preference indicators. We suggest that these more objective measures should be applied to enable the evaluation of image size and to formulate the guidelines with regard to image size settings. The importance of the subject is highlighted by the radiologists who participated in this study, of whom more than three-quarters indicated that they saw image size as a decisive factor in the diagnostic process. Though subjective data need to be interpreted with care, this suggests that the study of image size should be the focus of further research.

The present study has its limitations, which arise from the need to standardize the experimental setting. We chose two image sizes that are at the extremes of what can be displayed on the monitor that was employed. However, a majority of the participants indicated that they would prefer an image size between the two sizes presented in this study. It would therefore be interesting to include sizes between the two in a

Table 3 Reasons for radiologists' preference of a given image size.

	Preference for large images (N = 22)	Preference for small images (N = 16)	No preference for either image sizes (N = 5)
Motivation of preference			
More detail resolvable	15	0	0
Less tiring to read	3	2	0
Better contrast resolution	0	2	0
Better overview	0	10	0
More comfortable to read	1	0	0
Quicker to read	0	1	0
Size that I am used to	1	0	0
Small images for overview and large ones for detail	1	1	3
No reason given	1	0	2
I would have preferred an image size			
Smaller than the small images	0	1	0
Between the small and large images	8	9	3
Larger than the large ones	3	0	0
I perceive image size as a decisive factor in lesion detection			
Yes	18	11	4
No	4	5	1

future experiment. Furthermore, the experiment only looks at one modality and one type of lesion. It would be important to examine if these findings are also true for other settings. In a clinical setting, radiologists are able to adjust the image size during the process of interpretation. The participants of this study have been prevented from doing so in order to have the same form of presentation and the same image size for each participant. For similar reasons, no windowing was possible in this study. Participants have frequently complained about the lack of this option, and we conclude that the diagnostic accuracy, as measured by the JAFROC FOM, is lower than it would be in a more realistic setting. However, these limitations do apply to both image sizes and should therefore not question the validity of the conclusions of the experiment.

5 Conclusions

No connection between the preference for an image size and performance when reading cranial stack-mode CT was found in this experiment. This strongly suggests that the preference is not an indicator of performance when searching for intracranial hemorrhages in CT and should not be used as one. Instead, a combination of subjective measures, performance indicators, and methods that allow inferences by cognitive processes, such as eye tracking, could be used to understand more about the interpretive process.

Acknowledgments

This paper is a revised version of Venjakob, A.C., Marnitz, T., and Mello-Thoms, C. "Preference and performance regarding different image sizes when reading cranial CT" Proc. SPIE 9037, medical imaging 2014: Image perception, observer performance, and technology assessment, 903706 (11 March 2014). We would like to acknowledge and thank all 43 radiologists for their participation in the study. Further, we would like to thank Drs. Susan Grayson, Jean Mah and Kim-Son Nguyen, and Jan Mahler for their valuable help in collecting the data.

References

1. C. Schaefer et al., "Impact of hard-copy size on observer performance in digital chest radiography," *Radiology* **184**, 77–81 (1992).
2. S. E. Seltzer et al., "Influence of CT image size and format on accuracy of lung nodule detection," *Radiology* **206**(3), 617–622 (1998).
3. D. Gur et al., "The effect of image display size on observer performance: an assessment of variance components," *Acad. Radiol.* **13**, 409–413 (2006).
4. M. Yamaguchi et al., "Investigation of optimal viewing size for detecting nodular ground-glass opacity on high-resolution computed tomography with cine-mode display," *Radiol. Phys. Technol.* **4**(1), 13–18 (2011).
5. Y. Bessho et al., "Usefulness of reduced image display size in softcopy reading: evaluation of lung nodules in chest screening," *Acad. Radiol.* **16**, 940–946 (2009).
6. E. A. Krupinski "Practical applications of perceptual research," Chapter 20 in *Handbook of Medical Imaging. Volume 1: Physics and*

Psychophysics, R. L. Van Metter, J. Beutel, and H. L. Kundel, Eds., pp. 895–929, SPIE Press, Bellingham, Washington (2000).

7. S. R. Marschner and J. Lobb, “An evaluation of reconstruction filters for volume rendering,” in *VIS '94, Proc. Conf. Visualization '94*, IEEE Comput. Soc. Press, Los Alamitos, California (1994).
8. D. Chakraborty, JAFROC v 4.0, www.devchakraborty.com (20 February 2012)

Antje C. Venjakob studied psychology as an undergraduate and human factors as a master's degree. She recently completed her PhD thesis on visual search in medical multislice images and works as a research associate at Technische Universität Berlin, Germany.

Tim Marnitz is a clinical radiologist and a research associate in radiology at Charité Universitätsmedizin Berlin, Germany.

Lavier Gomes is a clinical associate professor of radiology at the University of Sydney. His research interests revolve around neuroimaging.

Claudia R. Mello-Thoms is an associate professor of medical radiation sciences at the University of Sydney and an adjunct professor at the University of Pittsburgh School of Medicine. Her research interests are in image perception, visual search, image interpretation, and cognitive modeling of medical decision making.