Title: Image size influences visual search and perception of hemorrhages when reading cranial CT - an eye tracking study

Authors: Antje C. Venjakob (M.Sc.)[a], Tim Marnitz (PhD)[b], Peter Phillips (PhD)[c], Claudia R. Mello-Thoms (PhD)[d]

Affiliations: [a]Technische Universität Berlin, Department of Psychology and Ergonomics, Chair of Human Machine Systems, Sekr. MAR 3-1, Marchstraße 23, 10587 Berlin, Germany

[b]Charité Universitätsmedizin Berlin - Campus Virchow Klinikum, Klinik für Radiologie Berlin, Germany

[c]University of Cumbria, Faculty of Health and Science, Lancaster, United Kingdom

[d]University of Sydney, Faculty of Health Science, Sydney, Australia & University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, United States of America

Running head: Effects of image size in cranial CT

Type of manuscript: Original investigation

Word count: 4885

Acknowledgments: We would like to thank all participating radiologists. Any correspondence regarding this article should be addressed to Antje Venjakob (antje.venjakob@mms.tu-berlin.de).

Abstract

Objectives: To explore reader gaze, performance and preference during interpretation of cranial computed tomography (cCT) in stack mode at two different sizes.

Background: Digital display of medical images allows for the manipulation of many imaging factors, like image size, by the radiologists, yet it is often not known what display parameters better suit human perception.

Materials and Methods: Twenty-one radiologists provided informed consent to be eye tracked while reading 20 cCT cases. Half of these cases were presented at a size of 14x14 cm (512x512 pixels), half at 28x28 cm (1024x1024 pixels). Visual search, performance and preference for the two image sizes were assessed.

Results: When reading small images significantly fewer, but longer fixations were observed, and these covered significantly more slices. Time to first fixation of True Positive findings was faster in small images, but dwell time on true findings was longer. Readers made more False Positive decisions in small images, but no overall difference in either JAFROC or reading time was found.

Conclusions: Overall performance is not affected by image size. However, small stack mode cCT images may better support the use of motion perception and acquiring an overview, whereas large stack mode cCT images seem better suited for detailed analyses.

Application: Subjective and eye tracking data suggest that image size influences how images are searched and that different search strategies might be beneficial under different circumstances.

Keywords: Radiology and Medical Imaging; Eye movements, tracking; Visual search; Computer interface

Précis: This study compares how radiologists interpret large versus small radiological stack mode images. No general performance differences were found for image size. Nonetheless, image size affected visual search, with patterns of gaze behavior usually associated with motion perception more likely for small images than for large images.

Introduction

Over the last few decades the work environment of radiologists has changed dramatically. Technologies like computed tomography (CT) and magnetic resonance imaging (MRI) have often replaced reading of single slice images like conventional radiographs. Furthermore, radiological imaging has gone digital, replacing the use of light-boxes by monitors (Andriole et al., 2011). With the increased use of digital reading, the number of factors that can be adjusted by the individual radiologist has increased enormously. One such factor is image size: it can be easily adjusted, and radiologists can set it according to their preference.

The empirical evidence with regard to the influence of image size on performance is mixed: several studies that used tile mode computed tomography, meaning that slices were presented next to each other on a light box (Schaefer at al., 1992; Seltzer et al., 1998), or single radiographs (Bessho, Yamaguchi, Fujita, & Azum, 2009) found disadvantages for small images (12 x 12 cm to 15 x 15 cm) compared to large images (30 x 30 cm), as measured by the area under the receiver operating characteristic (ROC) curve. However, two studies that aimed to ascertain performance relative to image size differences have found slight advantages for small images (Gur et al., 2006; Yamaguchi et al., 2011). In these experiments, stack mode CT images were used, meaning that slices were presented individually on the monitor and radiologists scrolled through them at their own pace. It can be hypothesized that the advantages for small images in these studies may be related to the form of presentation, as stack mode CT is a dynamic form of presentation that allows for the detection of lesions by motion sensitivity (Andia et al., 2009). Motion sensitivity can be exploited best by scrolling through the stack while resting the gaze in one position, rather than by scanning each image by multiple fixations. Indeed, recent eye tracking research has shown that 19 radiologists out of a sample of 24 rested their gaze in one position and scrolled through the stack, rather than searched each slice individually. These radiologists were termed ‘drillers’ to describe their preferred reading strategy. The performance of these participants was superior to that of the five radiologists who searched each slice individually. These participants were named ‘scanners’ (Drew, et al., 2013). As motion sensitivity is best towards the fovea (McKee & Nakayama, 1984; Pointer & Hess, 1989), this process should favor small images because these are covered more extensively by foveal and parafoveal vision. Additionally, motion sensitivity can be combined with better resolution in small images because these images can mostly be covered by high-resolution central vision. This primary central coverage would

potentially allow for a better discrimination between normal and abnormal structures, and thus help to flag out areas that need further visual inspection.

None of the previously conducted experiments on the effect of image size recorded eye position data, and hence the impact of images of different size on visual search during the interpretative process is largely unknown. Eye tracking studies have substantially contributed to our knowledge about perceptual and cognitive processes in medical imaging tasks, with many studies using it to gain insight into different sources of error (Kundel, Nodine, & Carmody, 1978), to study the time course of lesion detection (Mello-Thoms et al., 2005) or the layout of workstations (Krupinski, Roehring, & Furukawa, 1999). However, not many studies have yet used eye tracking to study perception and cognition in the context of volumetric medical images. This may partly be due to a more complicated setting where fixations span several slices and the calculation of classical parameters such as the time to first fixation having to be started from the first point in time when a lesion is visible rather than from the case onset (Phillips et al., 2013). Studies that have used eye tracking in the volumetric imaging context have so far often avoided these challenges by only using single images from the entire stack (Matsumoto et al., 2011) or by using raw data instead of fixation data (Drew et al., 2013; Drew, Vo, & Wolfe, 2013a). This can, however, complicate inferences about perceptual processes, such as the use of motion detection.

The present study takes a first step towards studying perceptual and cognitive processes in the interpretation of volumetric data by measuring and analyzing eye movements of radiologists across different slices when interpreting two different image sizes in digital cranial computed tomography (cCT) case sets. The main aim of this experiment is to gain insight into radiologists' visual search, perception and performance when reading digital multi-slice images of different size in stack mode. Secondary to this, the image size preference of the readers will be evaluated.

Methods

Participants

Twenty-one radiologists participated in the study, six of them female. At the time of the data collection all participants were employed by the same University hospital and participated during their working hours. The degree of experience in reading cCT varied between four months and 34 years, with a

mean experience of 6.4 years (*SD*= 6.3 years). At the time of participation they had been working for between 20 minutes and just over nine hours (*M*= 4.5h, *SD*= 2.8h).

Apparatus and Material

For the experiment 20 complete cCT cases were selected, which were rated normal by three independent radiologists who did not participate as readers in this study. All cases were acquired for clinical purposes, employing a 16-row spiral CT scanner (Light Speed, General Electric, Fairfield, Connecticut, USA) with an occipitomental angulation as unenhanced sequential CT of the head. Each case contained between 25 and 32 slices of 5 mm thickness. The 512 x 512 pixel DICOM images were de-identified, the contrast was set to the brain window of 35+/-40 Hounsfield Units (HU). Slices from each case were exported as uncompressed 8-bit Portable Network Graphics images. Then 18 subtle hemorrhages were cut from more severe cases displaying multiple lesions, and pasted into ten initially healthy cases, which resulted in 10 normal and 10 abnormal cases. The number of hemorrhages per abnormal case varied between one and three: four cases contained one hemorrhage, four contained two hemorrhages, and two cases contained three hemorrhages. Four of the hemorrhages were only displayed on one slice, twelve hemorrhages spanned two slices, and two hemorrhages spanned three slices. The effect of the image manipulation was assessed by three radiologists, who did not participate in the study. These radiologists did not realize they were looking at inserted hemorrhages, which led us to conclude that the insertions were successful. Figure 1 shows an example of a case with an inserted hemorrhage, indicated by the arrow.

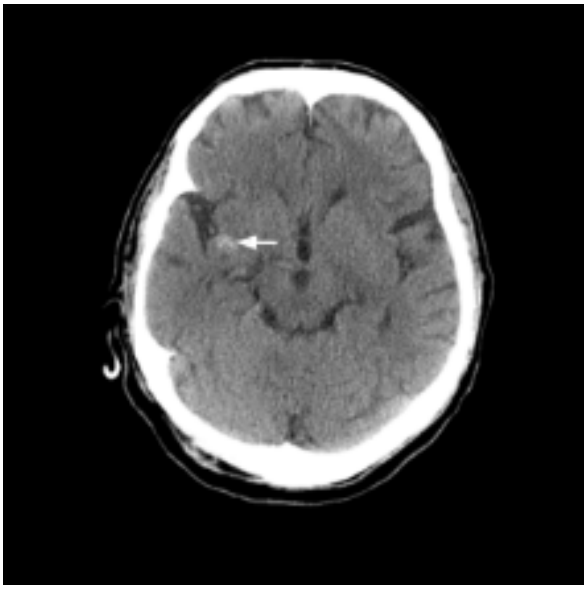

Figure 1. Slice with inserted intracranial hemorrhage indicated by an arrow.

All cases were saved in the size of 512 x 512 pixels then subsequently enlarged to a size of 1024 x 1024 pixels by employing the cubic Catmull-Rom-Splines interpolation algorithm (Marschner & Lobb, 1994) and saved in this size as well.

The images were inserted in MS PowerPoint 2010 slides. For each of the 20 cases two presentations were made: (i) a “small image size” presentation displaying 512 x 512 pixels images and (ii) a “large image size” presentation displaying 1024 x 1024 pixels images. As shown in Figure 2, the 20 cases per image size were divided into two sets of ten cases each. The lesion cases were equally distributed over the sets so that each image size always featured five lesion cases, altogether displaying nine hemorrhages.

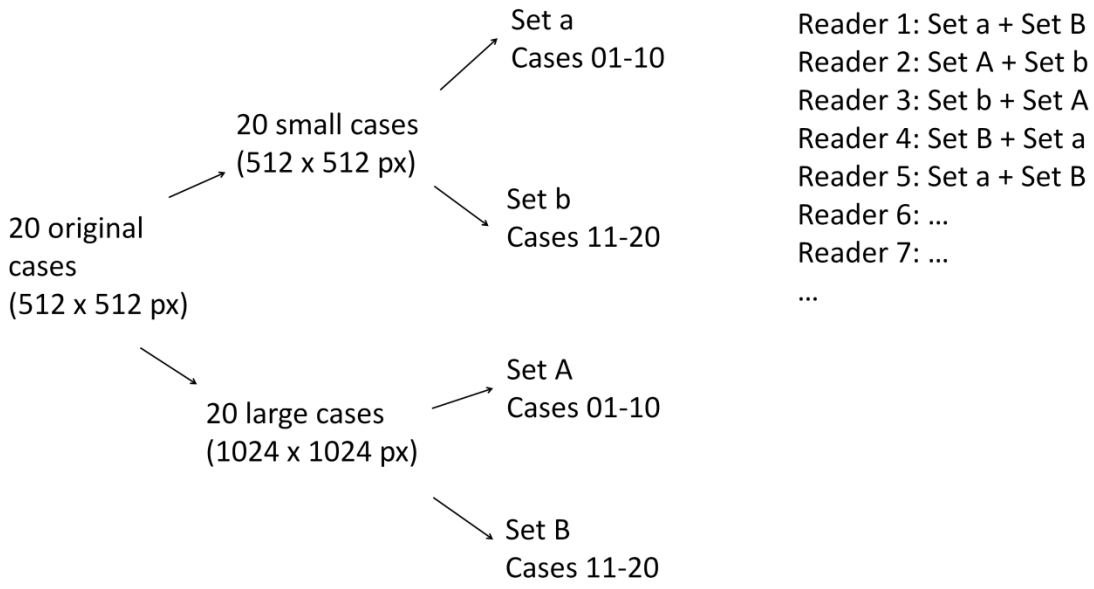

Figure 2. Display of large and small images in counterbalanced order between participants.

Only one set per image size was shown to a given radiologist, and no participant saw the same case twice. The presentation of the sets as well as their order was counterbalanced between readers. The order of the cCT cases of the same size was randomized within each set.

Slides were presented on a 1680 x 1050 pixels LCD DELL monitor connected to a remote eye tracker (SensoMotoric Instruments, iView X RED250 sampling at a frequency of 60Hz). The slide presentations and the eye tracking device were synchronized. No zooming or windowing was possible, but readers were able to scroll back and forth through the stack for as long as they wished to. A chin rest kept the viewing distance constant at 64 cm to prevent readers from compensating for smaller

images by approaching the monitor (Seltzer et al., 1998). Readers were instructed to use the mouse wheel to scroll though the slides and the left mouse button to encircle hemorrhages they chose to report along with a confidence rating on the presence of an hemorrhage, from 1 (very low confidence) to 10 (very high confidence). Figure 3 shows an example of a hemorrhage as reported by a reader.

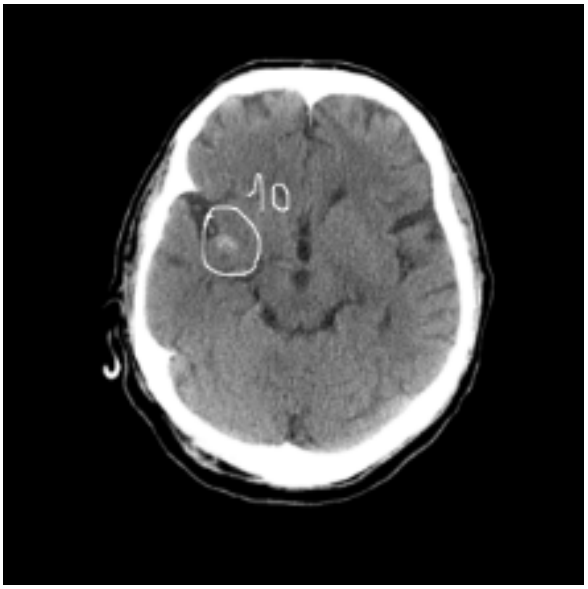

Figure 3. Inserted hemorrhage encircled and rated on a confidence scale of 1 to 10 by one of the participating radiologists.

A Java based tool was created to analyze the gaze data. It detects fixations, and matches these to the slices. Fixations were detected based on a low-speed dispersion algorithm. They were defined to be at least 80 milliseconds long, disperse no more than 2° visual angle in x and y direction and could span several slices. For fixations that spanned several slices, the relative dwell on each of the slices was calculated, i.e. for a fixation that spanned three slices, four values were calculated: the overall fixation duration, as well as the relative proportion of the fixation on each of the three slices. Subsequently, the tool calculated certain visual search parameters such as the time to first fixation on a given area of interest (AOI), dwell time per AOI and the number of slices covered by one fixation.

Experimental Design

The experiment consisted of a one-way within-subjects design. The within-factor was image size and had two levels: cranial CT case sets presented in 1024 x 1024 pixels (i.e., 28 x 28 cm), referred to as “large” images, and 512 x 512 pixels (14 x 14 cm), referred to as “small” images. Performance was quantified by the following dependent variables: the figure of merit (FOM) of the jackknife alternative free-response receiver operating characteristic (JAFROC), which is a variant of the ROC paradigm

that can take several decisions per case into account (Chakraborty, 2011), measured on the scale from 1(hemorrhage very unlikely) to 10 (hemorrhage very likely). Additionally, the number of True and False Positive and False Negative decisions per reader, and the median time to read a case was used to quantify performance. For the performance as well as eye tracking analysis a “True Positive” (TP) was scored whenever a participant encircled the location of an inserted hemorrhage, regardless of the confidence rating applied to it. A “False Negative” (FN) occurred when a radiologist failed to encircle one of the inserted hemorrhages. An encircled structure, that was not one of the 18 inserted hemorrhages, counted as the location of a “False Positive” (FP) decision. This was again independent of the confidence in the decision. For the eye tracking analysis additionally ten “True Negative” (TN) decision sites were selected for each reader. True Negative locations were chosen by randomly selecting five individual locations per reader and image size, where their gaze dwelled at least once but no hemorrhage was either present nor was one falsely reported. True Negative locations on average spanned 1.9 slices to align their size with that of True Positive and False Negative AOIs.

Visual search was quantified using the following per-case parameters: mean number of fixations, median fixation duration, and mean number of slices spanned by one fixation. Time to first fixation and dwell time were calculated per area of interest (AOI). In accordance with Phillips et al (2013), time to first fixation was defined as the time interval between the first appearance of an AOI and the start of the first fixation within it. Dwell time was the sum of the fixation durations on a given AOI over all slices that it appeared on, and over all visits to it during the reading of a case. Whenever a fixation fell partly into the AOI and partly outside it, e.g. because the reader scrolled onto a slice where the lesion is no longer visible, only the proportions of the fixation that fell into the AOI were included in the calculation of dwell time on the AOI.

All AOIs had the same size in small as well as in large images. The size was determined by the largest lesion plus a margin of 0.5° visual angle. This resulted in a radius of 1.5° visual angle (i.e. 65 pixels) around the center of each structure of interest. Equal AOI size for both image sizes was chosen because the visual field of the radiologists is equivalent in the two conditions. Hence, when the reader's dwell is located at 1.5° visual angle from the center of the lesion, both small and large lesions can be identified and avoided AOIs that are too small to account for eye tracker inaccuracy.

We sought to gain insight into perception by dividing errors of omission (i.e. the “False Negatives”) into search, recognition and decision making errors (Kundel, Nodine & Carmody, 1978). According to this

taxonomy, a search error is committed when a hemorrhage is not looked at, and recognition errors occur when an unreported hemorrhage is dwelled at for less than one second, thus preventing recognition that an abnormality is present at the location. Finally, decision making errors result from failure in correct identification, even though the hemorrhage was gazed at for more than one second.

Image size preference was assessed by asking radiologists to indicate, for each of the two modalities, how much they liked it on a continuous rating scale from zero (not at all) to ten (very much). They were further asked to make a binary choice by indicating if they preferred small or large images overall and to give a reason for their preference in free text. To gain insight into the relation between preference and performance, the continuous preference ratings of small and large images were correlated to the JAFROC measures of the respective image size.

Procedures

All readers were presented with an instruction screen, and completed a practice cCT before reading the two case blocks. At the beginning of each block a five-point calibration and four-point validation for the eye tracking system was performed. Where necessary, the system was recalibrated until an accuracy of less than 0.5° visual angle was achieved. At the end of the experiment a demographic questionnaire, and a questionnaire which assessed preference for the different image sizes, was filled out. The completion of the experiment took between 30 and 40 minutes.

Statistical Analysis

Normality of the data was tested using the Kolgomorov-Smirnoff test individually for each data subset and dependent variable. Where the normality assumption was met, the data was analyzed using paired sample t-tests, otherwise the non-parametric Wilcoxon signed-rank test for dependent samples was selected. Radiologists' preference ratings were correlated to their JAFROC performance using a Pearson correlation. The statistical analysis was performed in IBM SPSS Statistics version 21. For all tests, statistical significance was set at $p<0.05$.

Results

Visual Search

As shown in figure 4, the number of fixations per case in large images (*Mean*= 143.0, *SD*= 72.3) was greater than in small images (*M*= 109.0, *SD*= 49.4). This difference was statistically significant, $t(20)=$

3.4, *p*= 0.003. Figure 4 also shows that median fixation duration was significantly longer in small images (379 ms, *IQR*= 102) than in large images (304 ms, *IQR*= 175), *z*= 3.623, *p*< 0.001, and average fixations on small images covered significantly more slices than fixations on large images (*M*= 1.34, *SD*= 0.57; *M*= 1.05, *SD*= 0.52 respectively), *t*(20)= -3.83, *p*= 0.001.

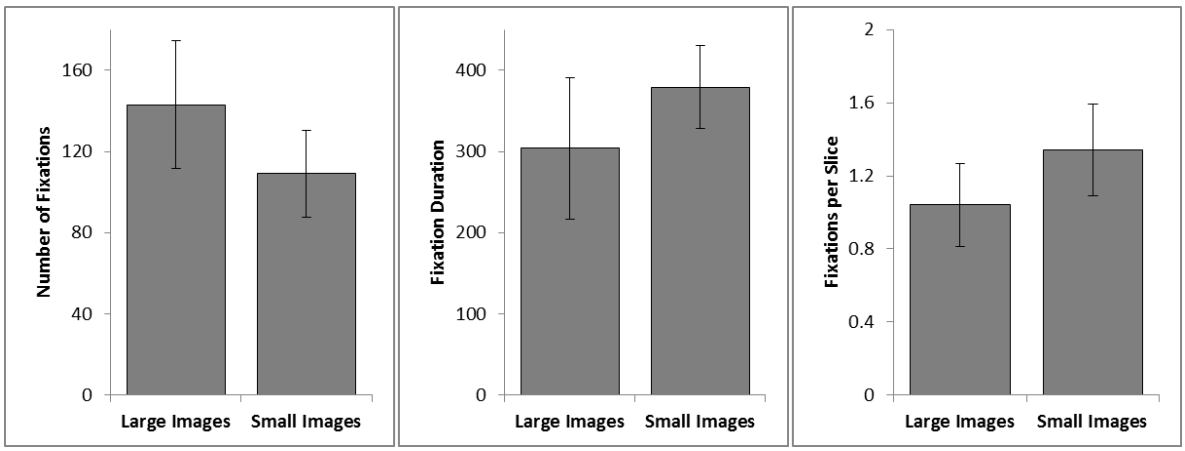

Figure 4. Mean number of fixations (error bar= standard error calculated based on the within-subject differences (Morey, 2008)), median fixation duration (error bar= interquartile range) and the mean number of slices that a fixation covered (error bar= Standard error calculated based on the within-subject differences (Morey, 2008)) are displayed separately for the two image size conditions.

Table 1 shows median values for dwell time and time to first fixation as well as their comparison using the Wilcoxon signed-rank test.

Table 1. Descriptive and inferential statistics of the parameters time to first fixation and dwell time for different decision outcomes and image sizes. Medians and interquartile ranges are shown in msec.

		Median Time to first fixation			Median dwell time		
		Mdn (IQR)	*Z*	*p*	*Mdn (IQR)*	*z*	*p*
True Positives	Large	937 (693)	-2.35	0.02	2626 (1992)	2.14	0.03
	Small	623 (331)			2619 (1785)		
True Negatives	Large	1002 (6850)	-1.03	0.30	662 (382)	2.9	0.01
	Small	835 (4312)			1204 (905)		
False Positives	Large	1712 (14621)	-1.65	0.10	3701 (3034)	0.47	0.64
	Small	1108 (4930)			3754 (2554)		
False Negatives	Large	3778 (-)	-1.60	0.11	2934 (-)	-1.60	0.11
	Small	666 (-)			483 (-)		

Time to first fixation was significantly shorter for True Positive locations in small images, with dwell time significantly longer on True Positive and True Negative locations. The significance test indicates longer dwell on True Positive locations in small images, even though the median is slightly smaller. This is due to the nature of the non-parametric tests, where absolute differences in the data are not taken into account, as significance is solely established by the ranking of the two alternatives.

The types of False Negative errors readers committed show different proportions of recognition and decision making errors for large and small images, as shown in Table 2. The number of decision-making errors was significantly lower for small as compared to large images (z= -2.6, p= 0.009), whereas small and large images did not differ significantly with regard to recognition and search errors.

Table 2. Different types of errors of omission for large and small cCT case sets. Number of errors committed shown, with percentages in brackets.

	Search errors	Recognition errors	Decision-making errors
Large	5 (25%)	3 (15%)	12 (60%)
Small	3 (25%)	8 (67%)	1 (8%)

Performance

In the large image condition, on average, 2.10 False Positives and 0.95 False Negative decisions were made, whereas in the small image condition there were 3.57 False Positive and 0.57 False Negative decisions on average. A within-subjects comparison using the Wilcoxon signed-rank test revealed that the difference with regard to the number of False Positive decisions was significant (z= 2.05, p= 0.04), whereas the difference in the number of False Negative decisions was not (z= -1.25, p= 0.21). Comparing the JAFROC scores of the two image size conditions yielded no significant difference (large images: *Median*= 0.68, *Interquartile range*= 0.11; small images: *Mdn*= 0.69, *IQR*= 0.07), z= 0.60, p= 0.95. Reading time was compared using the Wilcoxon signed-rank test for paired samples and there was no statistically significant difference between the image sizes (large images: *Mdn*= 55.1s, *IQR*= 35.3; small images: *Mdn*= 52.1 s, *IQR*= 37.2), z= -1.2, p= 0.23.

Image size preference

When making a binary choice, ten of the 21 participants preferred large images, eight preferred small images and three participants refused to choose between the two sizes and instead stated that they liked both. Preference on the continuous rating scale was on average 6.41 (*SD*= 2.07) for large images and 5.72 (*SD*= 2.46) for small images. This difference was not significant, $t(20)$= -0.84, p= 0.41, and neither preference ratings for small nor for large images correlated with the respective performance measures (large images: r= -0.15, N= 21, p= 0.52; small images: r= -0.07, N= 21, p= 0.75). Radiologists' motivation behind preference is displayed in Table 3.

Table 3: Motivation behind preference of a particular image size.

	Radiologists who prefer large images	Radiologists who prefer small images	Radiologists without preference
Motivation behind preference			
- More detail resolvable	7	0	0
- Less tiring to read	2	0	0
- Better contrast resolution	0	1	0
- Better overview	0	7	0
- Size that I am used to	1	0	0
- Small images for overview, large images for detail	0	0	2
- No reason indicated	0	0	1

Discussion

In this study we compared radiologists' visual search and perceptual processes when reading digital multi-slice images of different sizes presented in stack mode.

Changes in the perceptual processes when reading small- compared to large- images are best reflected in gaze behavior. When reading the smaller images readers made fewer fixations than when they read larger images, yet median fixation duration was longer for small images, and fixations span more slices. Together these findings suggest that when interpreting small images readers sought to take in information related to slice changes rather than by scanning the entire image. This hints to a more effective use of motion detection in small images, as the coverage of multiple slices by one fixation suggests that reading behavior can be compared to watching a movie rather than scanning a static image. In the terminology of Drew et al. (2013) this suggests that radiologists show more 'driller-like' behavior when reading small images compared to when reading large images, potentially suggesting that reading strategy is not only determined by an individual preference but also by environmental factors.

Using motion detection may be the preferred strategy with regard to small image because motion detection is better towards the fovea (Pointer & Hess, 1989): in this experiment the small images were about 14 cm wide, hence the entire image could be covered by a radius of 5° visual angle from the center of the fixation, often referred to as the useful field of view (UFOV). In the large images, the

farthest away areas are at 10° visual angle from the fixation center when the gaze rests in the middle of the image. At this distance motion detection has sharply deteriorated (McKee & Nakayama, 1984) as well as resolution (Carmody, Nodine & Kundel, 1980). Consequently, the combination of motion detection and improved resolution may be why True Positive findings are fixated, thus detected, faster in small than in large images and why radiologists who prefer small images indicate that these images provide a better overview.

However, these potential benefits do not translate to the overall performance data, as we observed that overall performance, measured by the JAFROC figure of merit and reading time, did not differ significantly between the image sizes. Small images show slightly better results regarding the JAFROC FOM and three seconds less are needed for their interpretation. The magnitude of the difference in the JAFROC FOM is roughly comparable to the difference found by Gur et al. (2004). However, Yamaguchi et al. (2011) found much greater, statistically significant differences between small and large images. Although the size of the images used in their study is comparable to the size of the images used in our experiment, a possible reason for this difference might be the use of different types of stimulus material. While Yamaguchi et al. used nodular ground-glass opacities, which tend to be subtle, our study employed hemorrhages which feature a rather bright contrast to neighboring tissue. For more subtle lesions, smaller images might be of greater advantage, as more tissue is covered by the UFOV in one fixation, while coverage by the UFOV may be of lesser importance when lesions are more conspicuous.

While overall performance did not differ between the two image sizes, the distribution of False Positive and False Negative decisions did change, as significantly more False Positive decisions were made in the small images. Conversely, on a descriptive level, more False Negative decisions were made in large images. Together, the findings seem to represent a shift in the decision criterion rather than a change in performance. Possibly, radiologists chose a more liberal decision criterion when interpreting smaller images. This might be a result of a loss of specificity, reflected in the higher number of False Positive decisions in small images and the finding that dwell time in small images is prolonged for True Positive and True Negative decision sites compared to dwell in these regions in large images. This may in turn be caused by a perceived decrease in image resolution in small images, which is suggested by the subjective data: radiologists who preferred large images said that these are better suited to resolve detail. This is the case although resolution is in fact not superior in large images in

this study, but is merely perceived to be so (Venjakob, Marnitz, Gomes, & Mello-Thoms, 2014). Two participants refused to choose between the two image sizes and instead indicated that they liked both for different reasons: small images for overview, large images for detail.

The comparison of motion detection processes between the image sizes by analyzing radiologists' fixations across slices was possible because entire cases were presented and explored by scrolling. The results hint to the necessity of studying entire cases rather than single slices of multi-slice cases. This study revealed that fixations, particularly in small images, tend to span multiple slices. This has an impact on fixation duration and presumably also on the distribution of dwell across an image or time to first fixation. It is, therefore, arguable whether the analysis of such parameters as it has been done in the past (Matsumoto et al., 2011) is representative of the clinical practice.

False Negative decisions seem to originate from different perceptual mechanisms in the two image sizes. The number of search errors did not differ between the two image sizes, which suggests that faster detection in small images does not automatically lead to the detection of more lesions. In small images, more recognition errors were committed, though this difference failed to reach statistical significance. Conversely, in large images, the high number of decision-making errors suggests that the unreported hemorrhages were identified as potentially abnormal sites, but faulty processing led the readers to dismiss the correctly detected abnormalities.

In our study many of the True Positive lesions were decided in less than one second, suggesting that the threshold to distinguish between recognition and decision errors potentially needs to be adjusted when used in the study of the interpretation of stack mode viewing. This might be the case because lesions could be more conspicuous due to the dynamic form of presentation. More data is needed to draw more definite conclusions regarding different types of False Negative errors in large and small images.

No correlation between preference and performance ratings was observed, suggesting that radiologists' subjective impression of useful parametrization might not necessarily be a valid indicator for beneficial reading conditions. This implies that guidelines for the interpretation process should not be based on preference ratings alone.

The study has several limitations that should be addressed: first, the insertion of lesions may have affected perception. Lesions were inserted with the utmost care and often spanned several slices

across which they gradually grew in size and then became smaller again to mimic the display of natural lesions. Nonetheless, there is always a possibility that edges and surrounding background tissue differed from what would normally be expected. This could have led to visual behavior different to what would usually be exhibited in the interpretation of hemorrhages. Secondly, the enlargement of the images may have led to a slight distortion in the image signal. Thus small and large variants of the same image might not have been exactly “identical” to one another. The cubic Catmull-Rom-Splines algorithm yields good results in terms of smoothing, but does so at the cost of less preferable results with regard to postaliasing (Marschner & Lobb, 1994). Nonetheless, no reader commented these issues, which leads us to believe that these effects, if present, were negligible. Further, in practice radiologists are free to adjust image size during the reading of a given case. It is possible that some readers use both the advantages of small images for detection and large images for decision-making, but have been prevented to do so in the laboratory setting of this experiment.

As the interquartile ranges suggest, there is a substantial amount of variance in the eye tracking data (see Table 1). This is particularly so for the parameter ‘time to first fixation’. However, we did not formally remove outliers from the data as they probably represent natural deviations in gaze behaviour when people perform a complex visual task on highly variable stimulus material.

Practical implications

No difference in global performance measures, as assessed by the JAFROC FOM and reading time, could be found, indicating that with regard to the interpretation of cranial CT the current practice of choosing an image size individually is not harmful. The findings are nonetheless of interest for day to day stack mode reading in the clinical practice as the subjective and eye tracking data of the study suggest that small images could be beneficial when perturbations are flagged and radiologists gain an overview over the case at hand. When more detailed information is needed for appropriate decision making, large images can be preferable to gain confidence in a decision and hence avoid a lack of specificity, which seems to be associated to small images.

Key points

- Different image sizes do not affect overall performance when radiologists read cranial CT.

- Eye tracking data suggests that image size affects visual search with more gaze behavior associated with motion perception being displayed in small as compared to large images.
- Subjective and eye tracking data suggest that image size influences how images are searched and that different search strategies might be beneficial under different circumstances.

References

Andia, M.E., Plett, J., Tejos, C., Guarini, M.W., Navarro, M.E., Razmilic, D, et al. (2009). Enhancement of visual perception with use of dynamic cues. *Radiology, 250,* 551-557.

Andriole, K.P., Wolfe, J.M., Khorasani, R., Treves, S.T., Getty, D.J., Jacobson, F.L., et al. (2011). Optimizing analysis, visualization, and navigation of large image data sets: One 5000-sectin CT scan can ruin your whole day. *Radiology*, *25,* 346-362.

Bessho, Y., Yamaguchi, M., Fujita, H., & Azuma, M. (2009). Usefulness of reduced image display size in Softcopy reading: Evaluation of lung nodules in chest screening. *Academic Radiology, 16*, 940-946.

Carmody, D., Nodine, C., & Kundel, H.(1980). An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception, 9,* 339-344.

Chakraborty, D.P. (2011). New developments in observer performance methodology in medical imaging. *Seminar of Nuclear Medicine, 41*, 401-418.

Drew, T., Vo, M. L.-H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision, 13*(10).

Drew, T., Vo, M. L.-H., & Wolfe, J.M. (2013a). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. Psychological Science, *24(9)*, 1848-1853

Gur, D., Klym, A., King, J., Maitz, G., Mello-Thoms, C., Rockette, H., & Thaete, F. (2006). The effect of image display size on observer performance: An assessment of variance components. *Academic Radiology, 13,* 409-413.

Kundel, H.L., Nodine, C.F., & Carmody, D.P. (1978). Visual scanning, pattern recognition and decision making in pulmonary nodule detection. *Investigative Radiology, 13,* 175-181.

Krupinski, E.A., Roehring, H., and Furukawa, T. (1999). Influence of film and monitor display

luminance on observer performance and visual search. *Academic Radiology, 6*, 411-418.

Marschner, S. R., & Lobb, R. J. (1994, October). An evaluation of reconstruction filters for volume rendering. In *Proceedings of the conference on Visualization'94* (pp. 100-107). IEEE Computer Society Press.

Matsumoto, H., Terao, Y., Yugeta, A., Fukuda, H., Emoto, M., Furubayashi, T, et al. (2011) Where do neuroradiologists look when viewing brain CT images? An eye tracking study involving stroke cases. *Plos one, 6: e282928*.

Mello-Thoms, C., Hardesty L., Sumkin, J., Ganott, M., Hakim, C., Britton, C., Stalder, J., & Maitz, G. (2005). Effects of lesion conspicuity on visual search in mammogram reading. Academic Radiology, 12, 830-840.

McKee, S.P. & Nakayama, K. (1984). The detection of motion in the peripheral visual field. *Vision Research, 24*, 25-32.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61-64.

Phillips, P., Boone, D., Mallett, S., Taylor, S., Altman, D., Manning, D., et al. (2013). Tracking eye gaze during interpretation of endoluminal 3D CT colonography: Technical description and proposed metrics for analysis. *Radiology, 267*, 924-931.

Pointer, J.S., & Hess, R.F. (1989). The contrast sensitivity gradient across the human visual field: with emphasis on the low spatial frequency range. *Vision Research, 29,* 1133-1151.

Schaefer, C., Prokop, M., Oestmann, J., Wiesmann, W., Haubitz, B., Meschede, A., et al. (1992). Impact of hard-copy size on observer performance in digital chest radiography. *Radiology, 184*, 77-81.

Seltzer, S.E., Judy, P.E., Feldman, U., Scarff, L., & Jacobson F. (1998). Influence of CT image size and format on accuracy of lung nodule detection. *Radiology, 206,* 618-622.

Venjakob, A. C., Marnitz, T., Gomes, L., & Mello-Thoms, C. R. (2014). Does preference influence performance when reading different sizes of cranial computed tomography?. *Journal of Medical Imaging, 1(3),* 035503-035503.

Yamaguchi, M., Bessho, Y., Inoue, T., Asai, Y., Matsumoto, T., & Murase K. (2011). Investigation of optimal viewing size for detecting nodular ground-glass opacity on high-resolution computed tomography with cine-mode display. *Radiological Physics and Technology, 4,* 13-18.

Biographies:

Antje Christine Venjakob studied Psychology as an undergraduate and Human Factors as a Masters' degree. She recently completed her PhD thesis on visual search in medical multi-slice images and works as a research associate at Technische Universität Berlin, Germany.

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

Peter Phillips is a computer scientist and obtained his PhD in medical image perception. He currently works as a lecturer and researcher at Cumbia University.

Claudia Mello-Thoms, PhD, is an Associate professor of Medical Radiation Sciences at the University of Sydney and an adjunct Professor at University of Pittsburgh School of Medicine. Her research interests are in image perception, visual search, image interpretation and cognitive modeling of medical decision making.