MIPS SPECIAL FEATURE

Multiple diagnostic task performance in CT examination of the chest

K M SCHARTZ, PhD, K S BERBAUM, PhD, M T MADSEN, PhD, B H THOMPSON, MD, B F MULLAN, MD, R T CALDWELL, MFA, B HAMMETT, MD, A N ELLINGSON, MD and E A FRANKEN JR, MD

Department of Radiology, The University of Iowa Roy J. and Lucille A. Carver College of Medicine, Iowa City, IA, USA

Objectives: In three experiments, we studied the detection of multiple abnormality types using the satisfaction of search (SOS) paradigm, the provision of a computer-aided detection (CAD) of pulmonary nodules and a focused nodule detection task.

Methods: 51 chest CT examinations (24 that demonstrated subtle pulmonary nodules and 27 that demonstrated no pulmonary nodules) were read by 15 radiology residents and fellows under two experimental conditions: (1) when there were no other abnormalities present except *test* abnormalities in the exams (non-SOS condition), and (2) when other abnormalities were present in the exams (SOS condition). Trials from the two conditions were intermixed. Readers were invited to return for two sessions: one in which the SOS condition was repeated with a simulated CAD; another in which only the non-SOS condition was presented. Detection accuracy was measured using receiver operating characteristic (ROC) analysis.

Results: An SOS effect (reduced detection accuracy for the *test* nodules in the presence of the diverse *added* abnormalities) was not found. Average accuracy was much higher when the CAD prompt was provided, without cost in the detection of the *added* abnormalities. Accuracy for detecting nodules appearing without intermixed SOS trials was also substantially improved.

Conclusions: CT interpretation was highly task dependent. Nodule detection was poor in the general search task. Therefore, CAD may offer a greater performance improvement than demonstrated in experiments assessing CAD using focused search. The absence of SOS may be due to limited nodule detection even without other abnormalities.

Advances in knowledge: CAD prompts of nodules increase the detection accuracy of nodules and decrease the time to detection—without impairing the detection accuracy—of non-nodule abnormalities.

Received 27 October 2011

Revised 17 April 2012

Accepted 26 April 2012

DOI: 10.1259/bjr/18244135

© 2013 The British Institute of Radiology

Modalities that provide multidimensional image data are now standard practice. The large number of images included in a single imaging examination may represent increased risk for failure of human perception [1, 2]. As imaging technology changes, what we consider a diagnostic error also changes. The capability to find smaller and subtler lesions dramatically increases the volume of images that need to be reviewed. There are so many image data that must be inspected that an exhaustive search may not be possible in any practical time frame. We know relatively little about perception in advanced medical imaging with few studies of how observers cope with these enormous data sets. With increasing information available, human attention becomes the limiting factor in diagnostic performance. The

observer's task—the nature of abnormalities sought and diagnostic questions addressed—determines how far his/her attention will be stretched. In the current experiments, we study the effects of observer task on error in multidimensional images.

This research begins with a failed experiment. Missing a second abnormality once another abnormality has been found is often called a “satisfaction of search” (SOS) error. Laboratory experiments provide an operational definition of SOS that cannot be found in retrospective accounts of errors. In this definition, the lesion that is missed because of SOS is shown to be detected in the absence of other lesions. But laboratory study of SOS effects has been limited to traditional radiography [3]. In fact, a high frequency of SOS errors has been hypothesised in retrospective studies of errors at CT [4]. Our original goal was to determine whether SOS effects occur in CT of the chest.

We compared the detection accuracy for simulated, but realistic “possibly malignant” pulmonary nodules in the presence and the absence of various naturally

Address correspondence to: Dr Kevin Schartz, Department of Radiology, 3174 Medical Laboratories, The University of Iowa, Iowa City, IA 52242, USA. E-mail: kevin-schartz@uiowa.edu
Supported by USPHS Grants R01 EB 00145 and R01 EB 00863 from the National Institute of Biomedical Imaging and Bioengineering, Bethesda, MD.

occurring abnormalities other than pulmonary nodules. In the language of the SOS paradigm, we compared the detection accuracy for *test* abnormalities in the presence and the absence of *added* abnormalities. Each "*test* abnormality" was a lung mass suspicious for malignancy. Each "*added* abnormality" was either a pulmonary abnormality other than lung mass or an abnormality of another organ visible in the CT examination.

This experiment was made possible when more complete control of the presence of both nodules and diverse native abnormalities became available by the creation of a software tool to remove, store and add abnormalities to medical images [5]. The simulated nodules inserted by this tool were designed to mimic primary lung malignancies, recognising that in many instances, particularly small nodules, distinctions between malignant and benign lesions may be difficult or impossible. These *test* abnormalities appeared twice, once with and once without another *added* abnormality. These *added* abnormalities were diverse, naturally occurring abnormalities that could be digitally removed. For examinations with *test* abnormalities, *added* abnormalities were removed from one version and retained in another. The *test* abnormality appeared in identical locations on both examinations (Figure 1). Thus, detection of the identical *test* abnormalities (pulmonary nodules) was measured in CT studies that were identical except for the presence of *added* abnormalities in the SOS condition. Our original null hypothesis was that the presence of *added* abnormalities would not produce any significant reduction in detection of *test* abnormalities (no SOS effect).

Unfortunately, detection of *test* pulmonary nodules was poor even without adding diverse abnormalities to CT examinations. Such low performance allows little opportunity to measure reduced performance indicating an SOS effect. We selected examples of subtle nodules for use as *test* abnormalities in the same way we had for many experiments with radiographs, based on the judgement of the subspecialist radiologists. All authors felt that the nodules would be found by most readers. But the nodules were too difficult for our observers, who were senior radiology residents.

How could the nodules seem obvious to experts, but be nearly invisible to near experts? Judging the salience

of an abnormality from the few CT slices containing it may fail to consider the navigation required to arrive at those few slices. We added a third experimental condition to study whether the nodules were actually visible to our observers when presented under the same conditions of interpretation as the experts who had selected the nodules. We provided simulated computer-aided detection (CAD) to discover whether the nodules were actually visible to observers if they were directed to them. The SOS condition was repeated as a single block of trials with CAD. Our second null hypothesis was that the *test* nodules would not be reported more accurately when localised by CAD than they had been in the original SOS condition.

Another possible reason for the poor performance of the observers is that the subspecialists looked only for nodules when they selected them for the experiment, whereas the observer's task was a general search. Both the experimental instructions to observers and the mixing of non-SOS and SOS trials in experimental sessions defined all abnormalities as fair game. Perhaps the general task definition itself reduces perceptual resources so far in CT interpretation that there is not much more decrement in performance than finding an abnormality can do on finding subsequent abnormalities. To study this possibility, we brought readers back for a fourth experimental condition. In this second follow-up experiment, the non-SOS condition was repeated as a single block of trials. Every non-nodule abnormality was eliminated from the task definition. Our third null hypothesis was that the *test* nodules appearing without any additional abnormalities would not be reported any more accurately in the "nodule-only" search task than they had been in the "general" search task.

Materials and methods

CT examinations

This research involved 51 CT examinations of the chest, 24 (47%) of which contained a pulmonary nodule serving as the *test* abnormality and 27 (53%) had no nodule present. Each case was seen with and without an *added* abnormality present.

Figure 1. The detection of a subtle pulmonary nodule was measured in the presence of a prominent abnormality—satisfaction of search (SOS) condition—and in its absence—non-SOS condition. A pulmonary embolus (white arrow) is apparent in one condition (a) and was digitally removed in a second condition (b). The detection of the pulmonary nodule (white arrow) as seen in (c) was measured in both conditions. The pulmonary embolus is most prominent with window and level pre-set for mediastinum and the pulmonary nodule is best seen with window and level pre-set for lung.

CT examinations were obtained from clinical studies with approval by our local institutional review board. Verification of the lesions and disease states was through follow-up studies, surgery, clinical course, laboratory tests and autopsy reports that were part of the patient medical record. All patient identifiers were removed from the CT studies to ensure patient confidentiality. The chest CT scans used in this research were performed on Aquilion (Toshiba, Otawara, Japan) and Sensation 16 (Siemens Medical Solutions, Forchheim, Germany) multislice scanners using standard helical scanning parameters: 135 kVp, 250–260 mAs, slice width of 3 mm and a pitch of 1.125. The CT field of view encompassed the smallest diameter of chest wall that completely contained the lung parenchyma as measured from widest point of outer rib to outer rib, typically 32–36 cm. Images were reconstructed to give contiguous, or overlapping, 3-mm-thick slices.

Abnormalities included in this experiment were selected by two chest radiologists (BHT, BFM) and two general radiologists (ANE, BH). The 24 *added* abnormalities that appeared with *test* nodules and the 27 *added* abnormalities that appeared without *test* nodules are delineated in (Table 1).

Image display

Examinations were presented on a CT workstation consisting of a Precision 360 Mini-tower (Dell, Round Rock, TX) and two 3-megapixel liquid crystal display monitors (National Display Systems, San Jose, CA) matched and calibrated to the Digital Imaging and Communications in Medicine (DICOM) standard using manufacturer's specifications. The calibration was checked periodically throughout the course of the experiment and adjusted when necessary.

We used *Workstation* software [6] (<http://perception.radiology.uiowa.edu>) to present the CT examinations and collect observer responses reporting abnormality and confidence rating for the abnormality. *Workstation* software was developed from *ImageJ*, a public domain image-processing and analysis package available from the National Institutes of Health (<http://rsb.info.nih.gov/ij>).

Removing diverse abnormalities and inserting pulmonary nodules

A thorough description of the procedures used to remove lesions and add nodules has been previously reported by Madsen et al [5]. Briefly, software was developed using Interactive Data Language (IDL) to remove abnormalities from sequences of CT images, store them in a library and place them within other CT examinations. In removal, the region occupied by the abnormality is replaced by normal tissue values sampled from the area surrounding the abnormality. Software was developed that allowed an operator to insert nodules from the database at any desired location and slice sequence of a CT series.

Although in the language of this SOS paradigm, the diverse abnormalities are called "*added*" abnormalities,

Table 1. Experimental design

Experimental condition	Non-SOS	SOS	SOS with CAD	Nodule search
Examinations contain an added abnormality?	No	Yes	Yes	No
Examinations contain a test abnormality?	Yes	Yes	Yes	Yes
Patients	27	27	27	27
Detection of test abnormalities (empirical ROC)	27 examinations with test abnormalities	27 examinations with test abnormalities	27 examinations with test abnormalities	27 examinations with test abnormalities
Sensitivity at specificity=0.9	For each of 15 readers	For each of 15 readers	For each of 13 readers	For each of 8 readers
Partial area [0, 0.25]	For each of 15 readers	For each of 15 readers	For each of 13 readers	For each of 8 readers
Comparisons:	Non-SOS vs SOS (15 pairs)	SOS vs SOS with CAD (13 pairs)		
	SOS vs SOS with CAD (13 pairs)	Non-SOS vs nodule search (8 pairs)		
Detection of added abnormalities (AFROC)—51 added abnormalities/51 false-positive images	Non-SOS vs nodule search (8 pairs)			
Sensitivity at specificity=0.9	N/A	For each of 13 readers	For each of 13 readers	N/A
Partial area [0, 0.25]	N/A	For each of 13 readers	For each of 13 readers	N/A
Comparisons:	SOS vs CAD with SOS (13 pairs)			

AFROC, alternative free-response ROC; CAD, computer-aided detection; N/A, not applicable; ROC, receiver operating characteristic; SOS, satisfaction of search.

they were actually native to the original CT examinations. Removal of “*added*” native abnormalities was used to generate the non-SOS versions of the CT examinations containing *test* abnormalities (*i.e.* the non-SOS CT examinations contained the *test* pulmonary nodule but did not contain an *added* abnormality). In most cases successive application of the removal operation was required to remove the abnormality. The software seamlessly meshed the replaced area without leaving any visual cues and allowed the operator to control the position, size and orientation of the replacement regions.

Our recipe for creating pairs of examinations with test nodules was straightforward. We began with 24 CT examinations that contained the diverse native abnormalities listed on the left side of Table 1. We made a copy of each examination and removed the abnormality from the copy. We then inserted a pulmonary nodule into each original and copy at the same location. (We actually tried several different nodules with different sizes and contrasts for each examination, in all cases settling on the most salient nodule.) With the nodules inserted, the 24 original CT examinations were part of the SOS condition and the 24 copies were part of the non-SOS condition.

Matching CT examinations without test nodules

27 paired examinations without *test* abnormalities (nodules) were used to measure false-positive rates. One member of the pair contained a readily identified abnormality (listed on the right side of Table 2) and the other did not present abnormality. The members of each pair were subjectively matched by size, lung capacity, and general anatomy (Figure 2).

Observers

15 volunteer radiologists from our radiology department were recruited as observers. They included 1 faculty member, 4 fellows, and 10 second-, third- or fourth-year

residents. All observers were given and signed an informed consent document that had been approved by our institutional review board for human subject use.

Experimental conditions: satisfaction of search

The experimental design for all experimental conditions is shown in Table 1. The observers saw each case twice in sessions separated by 2.5 months. The initial experiment consisted of two sessions separated in time to reduce the likelihood that the study images and the responses to them would be remembered. The presentation of non-SOS and SOS trials were intermixed in order to counterbalance the effect of whether examinations are seen for the first or second time by ensuring that effects of repeated presentation are equal in both treatment conditions. Half of the cases presented in each session contained an *added* abnormality and half did not. Thus, in the course of the two sessions, each CT examination appeared twice, once with and once without an *added* abnormality. If the CT examination appeared with an *added* abnormality in the first session, it appeared without the *added* abnormality in the second session. Similarly, if the CT examination appeared without the *added* abnormality in the first session, it appeared with the *added* abnormality in the second session. Within each session, examinations were presented in a pseudorandom order. The first and second sessions were separated by 2.5 months.

Ancillary condition: computer-aided diagnosis

CAD was not developed to counteract the SOS effect. To our knowledge, evaluations of observer performance informed by CAD have never studied the potential impact of CAD on detection of abnormalities that are not the target of the CAD algorithm. Current CAD finds nodules in chest CT examinations missed by unaided readers, but also reports false-positive locations. We

Table 2. “*Added*” abnormalities appearing as the satisfaction of search condition

Appearing with the test nodules	Appearing without test nodules
Liver lesion (4 examples)	Liver lesion (2 examples)
Pulmonary embolism (3 examples)	Aortic dilation (2 examples)
Gallstone (2 examples)	Aortic dissection (2 examples)
Mediastinal mass (2 examples)	Aortic aneurysm (2 examples)
Mediastinal node	Gallstone (2 examples)
Renal mass	Thyroid lesion (2 examples)
Renal cyst	Hilar mass/adenopathy (2 examples)
Precarinal node	Mediastinal mass
Thyroid nodule	Pulmonary embolism
Aortic dissection	Renal cyst
Gallbladder abnormality	Renal mass
Pancreatic calcifications	Mass behind posterior 10th rib
Tracheal mass	Extensive right pneumothorax
Gallbladder region mass	Multinodular goiter
Upper quadrant cystic mass	Right subclavian artery behind oesophagus
Chest wall fluid/mass	Adrenal nodule
Abdominal hernia	Rib fracture
	Soft tissue mass abutting the pleural surface
	Pneumonia
	Hepatic cyst

Figure 2. CT from four different patients. "Added" diverse abnormalities are indicated by white arrows. The mediastinal mass (a) is not present in a subjectively matched case (b); likewise, the hepatic cyst (c) is not present in (d).

added this experimental condition to test whether CAD prompts for nodule detection would improve detection in the presence of *added* abnormalities. 13 of the original readers returned after 4–5 months to reread the examinations of the SOS condition presented as a single block in which a simulated CAD prompt indicated the location of the *test* nodules when they were present and of pulmonary vessels when *test* nodules were not present. The CAD prompt could be obtained by a button on the navigation menu. However, observers were instructed to inspect the study images unaided before accessing the computer aid. Part of our motivation for this subsequent experiment was to discover whether nodules serving as *test* abnormalities would be considered visible when pointed out by the CAD prompt.

We compared the performance in detecting the *added* abnormalities when they appeared in the SOS experiment with when they appeared with simulated CAD in the ancillary condition. For this comparison, we scored a response as true-positive if it was within the region of interest of an *added* abnormality. A response was scored as false-positive if it was outside the regions of *test* and *added* abnormalities, and the response description referred to features outside the lung, but did not refer to *added* abnormalities.

Ancillary condition: nodule search

We suspected that the procedure of the SOS experiment in which trials from different experimental conditions were intermixed in each experimental session may affect search strategies to a greater extent than we have observed in radiography. Seeing and reporting the diverse non-nodule abnormalities present in the SOS trials may affect how the reader interprets his/her task and in turn search on non-SOS trials. Therefore, we repeated the non-SOS trials (*test* nodules without *added* abnormalities) as a separate single block of trials with reading started in the lung window. 8 of the original 15

readers returned after approximately 18 months to participate in this experiment.

Procedure

Prior to the start of the experiment, each observer was read instructions and shown a demonstration case on how to display the images, make responses and advance through the cases. Observers were instructed to search for clinically significant abnormalities and to identify each abnormality by placing the mouse cursor over it and clicking with the right mouse button. This produced a text box for entering a brief description of the abnormality. They rated their confidence that the feature was abnormal. The observers were directed to indicate their confidence that a feature was abnormal by using a scale from 10 to 100% certainty, with 10% increments. Examinations without responses were scored as 0% certainty. For the non-SOS, SOS and CAD/SOS conditions, observers were instructed to begin search with the mediastinal window/level pre-set and afterward to continue searching using lung and bone window/level pre-sets. For the non-SOS/nodule search condition, observers were instructed to begin search for nodules using the lung window.

Receiver operating characteristic scoring of test abnormalities

Rectangular regions were identified around our abnormalities on each CT slice in which they appeared so that locations of responses could be classified as corresponding to (1) our *added* abnormalities, (2) our *test* abnormalities or (3) normal anatomy. On cases without *test* abnormalities, a response had to fall outside the regions of abnormalities to be scored false-positive. In scoring two types of abnormalities, we needed a way to associate each false-positive response with *added* or

test abnormalities. We could consider all false-positive responses on examinations without nodules towards the receiver operating characteristic (ROC) analysis of the *test* abnormalities (nodules). When we did so, we found an SOS reduction in accuracy that was based on a higher false-positive rate in the SOS condition. Many of the false-positive responses were not identified as nodules or even localised within the lung field. It seemed that the introduction of diverse abnormalities and responses to them brought along their own set of false-positive responses. The SOS obtained this way is unlike the SOS we ordinarily find which involves a reduction in true-positive reports of *test* abnormalities without change in false-positive reports. To achieve a more precise scoring of false-positive responses, we separated them into those falling inside and outside the lung. So, false-positive responses inside the lung (largely non-existent pulmonary nodules) were considered in the ROC analysis of nodule detection; false-positive responses outside the lung (largely reported in the mediastinal window) were considered in the ROC analysis of *added* abnormalities.

For each response, the observer typed a brief description of the abnormality being reported. We classified these descriptions without knowledge of whether they were made in the non-SOS or SOS conditions. Reported abnormality type was used for sorting false-positive responses. Responses where the description referred to the type of abnormality *added* to a particular examination were not considered false-positive. Descriptions were sorted into those referring to features inside the lung, including reports of pulmonary nodules, and those referring to features outside the lung. For each examination without a *test* abnormality, the one false-positive response with the highest confidence was scored.

We scored responses as true-positive that were within the *test* abnormality's region of interest. False-positive responses were outside the regions of *test* and *added* abnormalities, and had descriptions that did not refer to *added* abnormalities, but did refer to features inside the lung.

Receiver operating characteristic data analysis

We used the empirical ROC method to analyse our data. For each treatment-reader combination, we used the sensitivity at specificity of 0.9 [alternatively the true-positive fraction at the false-positive fraction (FPF) of 0.1] and the partial area under the empirical ROC curve for the FPF interval of [0.0, 0.25] (subsequently referred to as "partial area [0, 0.25]") as indices of accuracy. Both approaches avoid the problems associated with ROC area for empirical curves [7, 8]. Sensitivity at fixed specificity has the advantage of being readily understood by non-statisticians, whereas simulation studies have suggested that partial area may enjoy an advantage in statistical power [9]. Specificity of 0.9 places the measured sensitivities near the centre of available ROC points for detecting *test* nodules for all conditions for across readers. Experimental conditions were compared with non-parametric Wilcoxon signed-rank tests. These tests generalise to the population of radiologists/radiology residents rather than the population of patients [10].

Shift in decision thresholds

Shifts in decision thresholds are of special interest when radiological consultation is provided by human or machine consultants. Swensson et al [11–14] demonstrated that increases in true-positive rates balance with increases in false-positive rates when observers evaluated specific locations for pulmonary nodules that either contained a nodule or were the site of a previous false-positive nodule call by another radiologist. According to signal detection theory, balanced changes in true- and false-positives result from shifts in decision thresholds that bound the various rating categories [15]. Providing locations at which a prior reader reached a positive conclusion produced overcalling, a response bias that produces more true- and false-positive responses. This effect, first demonstrated by Swensson in 1977, years before the first CAD algorithms in radiology, should properly be called the "Swensson effect". It has been demonstrated more recently that limiting the maximum number of cued regions allowed per case improved the overall performance of CAD in mammography [16], providing a very practical demonstration that Swensson's conclusions about radiologic consultation were correct.

Analysis of decision thresholds is not very common in radiology research. One method is to compute the false-positive rates associated with decision thresholds estimated by fitting an ROC model [17, 18]. Because we rely on empirical ROC analysis in the current paper, analysis of subjective probability ratings unrelated to either *added* or *test* abnormalities for examinations without *test* abnormalities is a straightforward approach to gauging response bias. It should be noted that analyses of raw rating data for separate normal and abnormal cases is not new in medical image perception research [19, 20]. The practice can be traced to Swensson, who would often insist on these supplemental analyses to better understand the FPF and true-positive fraction (TPF) changes underlying ROC curves.

Response time

We computed median response times within each treatment condition. The means of the reader medians for different treatments were compared using non-parametric Wilcoxon tests. The means for these tests and the associated degrees of freedom reflect the requirement that reader medians be available for both conditions being compared.

Results

Detection analysis

ROC points and empirical curves for detection of the test nodules from all readers are presented in Figure 3. The TPF, sensitivity, is plotted against the FPF, the complement of specificity. The non-SOS condition is indicated by solid squares with solid lines, the SOS condition is indicated by open triangles with solid lines, the SOS with CAD condition is indicated by open circles

Figure 3. The receiver operating characteristic (ROC) points and empirical ROC curves for detection of *test* nodules for 15 readers in the satisfaction of search (SOS) condition (open triangles with solid lines), the non-SOS condition (filled squares with solid lines), the SOS condition with computer-aided detection condition (open circles with dashed lines) and the non-SOS/nodule search condition (open squares with dashed lines). The true-positive fraction (TPF or sensitivity) is plotted against the false-positive fraction (FPF or the complement of specificity).

and dashed lines, and the nodule search condition is indicated by open squares and dashed lines. Readers 7 and 14 did not participate in the CAD condition, and Readers 1, 2, 4, 5, 6, 8 and 14 did not participate in the nodule search condition.

For many readers the ROC curves for the non-SOS condition lie at chance (Readers 3 and 9) or near chance (Readers 4, 6, 10, 12, 13, and 14). While other readers had non-SOS performance above chance, the curves were still fairly low. For those readers participating in the CAD condition, ROC curves were frequently quite high. For those readers participating in the nodule search condition, ROC curves were between the two extremes.

In the presentation of the results, each mean is provided with its standard error in parentheses. Mean differences and standard errors of the mean differences are provided for statistically significant results.

Added abnormalities did not produce a satisfaction of search effect for detection of *test* abnormalities. The average empirical ROC sensitivity at specificity=0.9 for detecting *test* nodules for the non-SOS condition (without *added* abnormalities) and the SOS condition (with *added* abnormalities) was not statistically significant: 0.33 (± 0.04) *vs* 0.32 (± 0.03), Wilcoxon=40.0 (15 pairs), $p=0.70$. The average partial area [0, 0.25] for detecting *test* nodules for the non-SOS condition and the SOS condition was not statistically significant: 0.08 (± 0.01) *vs* 0.08 (± 0.01), Wilcoxon=33.0 (15 pairs), $p=0.64$.

Providing CAD to the SOS condition greatly enhanced detection of *test* abnormalities relative to the main experiment SOS condition without CAD. Average ROC sensitivity at specificity=0.9 was much higher when the CAD prompt was added: 0.79 (± 0.07) *vs* 0.32 (± 0.04), mean difference=0.47 (± 0.07), Wilcoxon=1.0 (13 pairs), $p=0.0005$. Average partial ROC area was much higher when the CAD prompt was added: 0.08 (± 0.01) *vs* 0.19 (± 0.01), mean difference=0.11 (± 0.01), Wilcoxon=0.0 (13 pairs), $p=0.0002$. Of course, the simulated CAD had very high performance: perfect sensitivity and only one false-positive response for examinations without nodules. Thus, our observers agree that the nodules were reportable when directed to them.

Test abnormality detection was enhanced for the non-SOS condition when search focused on nodule detection and none of the trials in the session presented *added* abnormalities. Average ROC sensitivity at specificity=0.9 for detecting the *test* abnormalities was significantly higher in the non-SOS trials from the single block in the ancillary experiment than in the non-SOS interleaved trials from the main experiment: 0.54 (± 0.05) *vs* 0.34 (± 0.08), mean difference=0.20 (± 0.07), Wilcoxon=1.0 (8 pairs), $p=0.016$. There was a nearly significant finding for average partial ROC area for detecting the *test* abnormalities being higher in the non-SOS trials from the single block in the ancillary experiment than in the non-SOS interleaved trials from the main experiment: 0.13 (± 0.01) *vs* 0.08 (± 0.02), mean difference=0.04 (± 0.02), Wilcoxon=4.0 (8 pairs),

$p=0.055$. Thus, the non-SOS trials may have been affected by the presence of the SOS trials that were intermixed with them.

Native abnormality detection for the satisfaction of search condition with and without computer-aided detection

Separate empirical ROC analyses of ratings in *added* abnormality regions and normal regions in those same examinations (with *added* abnormalities) were performed. This type of scoring is called alternative free-response ROC (AFROC) scoring. ROC points and empirical curves for detection of the *added* abnormalities from 13 readers are presented in Figure 4. The TPF, sensitivity, is plotted against the FPF, the complement of specificity. The SOS condition without CAD is indicated by open triangles with solid lines, and the SOS with CAD condition is indicated by open circles and dashed lines. No systematic difference between these two conditions across the observers is apparent in the figure. Providing CAD to the SOS condition neither enhanced nor hurt detection of *added* abnormalities relative to the SOS condition without CAD. Average ROC sensitivity at specificity=0.9 was no different when the CAD prompt was added: 0.52 (± 0.05) *vs* 0.47 (± 0.05), mean difference=0.05 (± 0.05), Wilcoxon=30.0 (13 pairs), $p=0.28$. Average partial ROC area was no different when the CAD prompt was added: 0.13 (± 0.01) *vs* 0.12 (± 0.01), mean difference=0.01 (± 0.01), Wilcoxon=32.0 (13 pairs), $p=0.35$.

Decision threshold shifts with computer-aided detection

Analysis of subjective probability percentage ratings for examinations without *test* nodules (false-positive ratings) can reveal shifts in response probabilities. The ratings for the 27 examinations with *test* nodules absent were averaged for each treatment condition. The resulting pairs of averages (for the SOS condition without CAD and the SOS condition with CAD) were compared using a Wilcoxon signed-rank test. There was a significant increase in the percentage certainty of *test* nodules or other lung abnormality with CAD over the SOS condition (5% for SOS without CAD *vs* 23% for SOS condition with CAD, $p=0.0002$).

Response time

Nodule search

Test abnormalities (nodules) were reported relatively late in search regardless of whether native abnormalities were present or CAD prompts were provided. However, *test* abnormalities were reported much earlier in the

Figure 4. The receiver operating characteristic (ROC) points and empirical ROC curves for detection of *added* abnormalities for 13 readers in satisfaction of search (SOS) condition (open triangles with solid lines), and the SOS condition with computer-aided detection condition (open circles with dashed lines). The true-positive fraction (TPF or sensitivity) is plotted against the false-positive fraction (FPF or the complement of specificity).

second ancillary experiment where search was directed at nodules. Results have been rounded to the nearest second.

The time to report *test* abnormalities (pulmonary nodules) did not differ for the non-SOS, SOS, and SOS with CAD conditions, but was shorter for non-SOS as a block of trials. The average median time to report *test* nodules did not differ for the non-SOS and SOS conditions of the main experiment: 111 (± 17) vs 114 (± 16) seconds, Wilcoxon=33.0 (13 pairs), $p=0.38$. The average median time to report *test* nodules did not differ for the SOS condition of the main experiment and the SOS condition with CAD from the ancillary experiment: 116 (± 18) vs 129 (± 12) seconds, Wilcoxon=19.0 (11 pairs), $p=0.21$. The average median time to report *test* nodules was significantly shorter for the non-SOS condition in a block of trials from the ancillary experiment than for the non-SOS condition of the main experiment: 28 (± 4) vs 90 (± 13) seconds, mean difference=61 (± 10) seconds, Wilcoxon=0.0 (6 pairs), $p=0.031$.

Search for added abnormalities

When *added* abnormalities were present, the time to report them was less than the time to report *test* abnormalities. For the SOS condition, true-positive *added* abnormality reports occurred significantly earlier than true-positive *test* abnormality reports: 58 (± 13) seconds vs 114 (± 16) seconds, mean difference =57 (± 17) seconds, Wilcoxon=11.0 (13 pairs), $p=0.0159$. For the SOS plus CAD condition, true-positive *added* abnormality reports occurred significantly earlier than true-positive *test* reports: 24 (± 3) seconds vs 120 (± 12) seconds, mean difference=96 (± 10) seconds, Wilcoxon=0.0 (13 pairs), $p=0.0002$.

The average of median time to report *added* abnormalities was significantly shorter when CAD was provided than for the SOS condition of the main experiment: 24 (± 3) vs 55 (± 13) seconds, mean difference=31 (± 12) seconds, Wilcoxon=0.0 (13 pairs), $p=0.0112$.

Discussion

Perception of advanced imaging has received very little study. Interpretation of today's CT examinations is quite different from radiographs. Detection tasks are generally nested within window width and level settings specific to lung, mediastinum and bone, and much more time is required. The image data are not all available at once and only revealed as navigation progresses through the stack of images. Successive presentation of the reconstructed sections involves the motion-detection systems of the brain. This more serial inspection may render interpretation of CT more dependent on the reader's task definition.

This research began with a failed experiment on SOS in chest CT. The *test* nodules in the original non-SOS/

SOS comparison were too hard for readers to detect even without *added* abnormalities. The overall low detection of *test* abnormalities did provide unanticipated lessons: judging the salience of abnormalities by inspecting the few CT slices containing them fails to take into account the navigation required to arrive at those slices. In addition, using a nodule search task may select nodules too difficult to find in a general search task. Unfortunately, these lessons leave the original question unanswered. Further research is needed to discover whether SOS effects can be demonstrated in chest CT with more detectable *test* abnormalities. Experiments are in progress in which many of the *test* nodules we used here are replaced with much larger ones and the mixing of experimental conditions with sessions is systematically studied. Before we started to see the results of these in-progress experiments, we could not believe that the replacement nodules, several times larger, would ever be missed. The original nodules were detected at rates of 10–60%, but the larger, enhanced nodules are still only being detected at rates of 50–85%.

In adding a condition prompting the locations of possible nodules and a condition limiting search to nodules, we hoped to discover whether the nodules were actually visible to our observers. Limiting search to nodules did improve nodule detection and prompting locations yielded high response rates and dramatically better detection. These findings did speak to the nature of the design problem in our original SOS experiment, but also provided new findings about single- and multiple-task performance in advanced imaging.

Our experiment was not the first to compare single- and multiple-task performance. In chest radiographs, Samuel et al [21] used pulmonary nodules as the *test* abnormalities and diverse abnormalities as *added* abnormalities creating a new SOS paradigm. The study demonstrated that detection of the *test* nodules was dependent on the observer's task. For radiographs without native abnormalities (the non-SOS condition), detection rate for nodules decreased from 72% for a nodule-directed task to 29% for a general search task. (This reduction was even greater when native abnormalities were present, demonstrating an SOS effect.) We wondered whether the low performance in detecting *test* abnormalities in the absence of *added* abnormalities in our experiment was due to this same task dependence. Perhaps the radiologists selecting our nodules operated under different task demands than our experimental readers.

Beyond the experimenter's instructions, laboratory readers understand their task on the basis of the abnormalities they find as they advance through the experiment. When SOS trials are present, intermixed with the trials of the non-SOS control condition, finding the *added* abnormalities would keep the reader's task general. Subsequent non-SOS trials could then be treated as a general search task. By contrast, the nodule-search condition presented only *test* nodules without *added* abnormalities (non-SOS) and began each examination

with the lung window. Consistent with the findings of Samuel et al [21], the nodule search task yielded a substantial improvement over the general search task.

More surprising results were obtained with the addition of CAD to the SOS condition. When nodules were pointed out, they became visible; nodule detection with CAD was dramatically better. Admittedly, our simulated CAD may offer less ambiguous help than real CAD algorithms. The performance advantage for detecting nodules from limiting the reader to a nodule detection task was small relative to that from CAD. Despite the large improvement in nodule detection, the more novel results derive from the multitask nature of these experiments: the effect of CAD on general search has not been studied. The enhancement in nodule detection from the CAD prompt occurred without reduction in detection of various other abnormalities. The CAD prompts had no effect on the amount of time needed to find the nodules, but reduced the time to find the other abnormalities. Perhaps readers remove the nodule-search task from their early general search, postponing search for the CAD-targeted abnormality until the CAD prompt was provided.

References

1. Armato SG, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. *Med Phys* 2001;28:1552–61.
2. Andriole KP, Wolfe JM, Khorasani R, Treves ST, Getty DJ, Jacobson FL, et al. Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day. *Radiology* 2011;259:346–62.
3. Berbaum KS, Franken EA Jr, Caldwell RT, Scharzt KM. Satisfaction of search in traditional radiographic imaging. In: Samei E, Krupinski E, Jacobson F, eds. *The handbook of medical image perception and techniques*. Cambridge, UK: Cambridge University Press; 2010. pp. 107–38.
4. White CS, Romney BM, Mason AC, Austin JHM, Miller B, Protopapas Z. Primary carcinoma of the lung overlooked at CT: analysis of findings in 14 patients. *Radiology* 1996;199:109–15.
5. Madsen MR, Berbaum KS, Caldwell RT. A new software tool for removing, storing and adding abnormalities to medical images for perception research studies. *Acad Radiol* 2006;13:305–12.
6. Scharzt KM, Berbaum KS, Caldwell RT, Madsen MT. WorkstationJ: workstation emulation software for medical image perception and technology evaluation research. In: Jiang Y, Sahiner B, eds. *Medical imaging 2007: image perception, observer performance, and technology assessment*. Proc SPIE 2007;6515:1–11.
7. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data. Part II: a formal model. *Acad Radiol* 2000;7:427–37.
8. Gur D, Bandos AI, Rockette HE. Comparing areas under receiver operating characteristic curves: potential impact of the “last” experimentally measured operating point. *Radiology* 2008;247:12–15.
9. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press; 2004. pp. 79, 224, 228.
10. Berbaum KS, Dorfman DD, Franken EA. Measuring observer performance by ROC analysis: Indications and complications. *Invest Radiol* 1989;24:228–33.
11. Swensson RG, Hessel SJ, Herman PG. Omissions in radiology: Faulty search or stringent reporting criteria? *Radiology* 1977;123:563–7.
12. Swensson RG, Hessel SJ, Herman PG. Radiographic interpretation with and without search: visual search aids the recognition of chest pathology. *Invest Radiol* 1982;17:145–51.
13. Swensson RG, Hessel SJ, Herman PG. The value of searching films without specific preconceptions. *Invest Radiol* 1985;20:100–14.
14. Swensson RG, Theodore GH. Search and nonsearch protocols for radiographic consultation. *Radiology* 1990;177:851–6.
15. Green DM, Swets JA. *Signal detection theory and psychophysics*. New York, NY: John Wiley and Sons; 1962. pp. 86–116.
16. Zheng B, Leader JK, Abrams G, Shindel B, Catullo V, Good WF, et al. Computer-aided detection schemes: the effect of limiting the number of cued regions in each case. *AJR Am J Roentgenol* 2004;181:579–83.
17. Franken EA, Berbaum KS, Lu CH, Kannam S, Dorfman DD, Warnock NG, et al. Satisfaction of search in detection of plain film abnormalities in abdominal contrast examinations. *Invest Radiol* 1994;29:403–9.
18. Berbaum KS, Franken EA Jr, Caldwell RT, Scharzt KM. Satisfaction of search in traditional radiographic imaging. In: Samei E, Krupinski E, Jacobson F, eds. *The handbook of medical image perception and techniques*. Cambridge, UK: Cambridge University Press; 2010. pp. 107–38.
19. Berbaum KS, Franken EA, Dorfman DD, Barloon T. Influence of clinical history upon detection of nodules and other lesions. *Invest Radiol* 1988;23:48–55.
20. Berbaum KS, Franken EA, Jr, Anderson KL, Dorfman DD, Erkonen WE, Farrar GP, et al. The influence of clinical history on visual search with single and multiple abnormalities. *Invest Radiol* 1993;28:191–201.
21. Samuel S, Kundel HL, Nodine CF, Toto LC. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. *Radiology* 1995;94:895–902.