Published in final edited form as:

Acad Radiol. 2013 February ; 20(2): 194–201. doi:10.1016/j.acra.2012.08.017.

Satisfaction of Search from Detection of Pulmonary Nodules in Computed Tomography of the Chest¹

Kevin S. Berbaum, PhD, Kevin M. Schartz, PhD, Robert T. Caldwell, MFA, Mark Madsen, PhD, Brad H. Thompson, MD, Brian F. Mullan, MD, Andrew N. Ellingson, MD, and Edmund A. Franken Jr., MD

¹Department of Radiology, The University of Iowa Roy J. and Lucille A. Carver College of Medicine, Iowa City, IA 52242

Abstract

Rationale and Objectives—We tested whether satisfaction of search (SOS) effects occur in computed tomography (CT) examination of the chest on detection of native abnormalities is produced by the addition of simulated pulmonary nodules.

Materials and Methods—Two experiments were conducted. In the first experiment, seventy CT examinations, half that demonstrated diverse, subtle abnormalities and half that demonstrated no native lesions, were read by 18 radiology residents and fellows under two experimental conditions: presented with and without pulmonary nodules. In a second experiment, many of the examinations were replaced to include more salient native abnormalities. This set was read by 14 additional radiology residents and fellows. In both experiments, detection of the naturally abnormalities was studied. ROC curve areas for each reader-treatment combination were estimated using empirical and proper ROC models. Additional analyses focused on decision thresholds and visual search time on abnormality-free CT slice ranges. Institutional review board approval and informed consent from 32 participants were obtained.

Results—Observers more often missed diverse native abnormalities when pulmonary nodules were added but also made fewer false positive responses. There was no change in ROC area, but decision criteria grew more conservative. The SOS effect on decision thresholds was accompanied by a reduction in search time on abnormality-free CT slice ranges.

Conclusion—The SOS effect in CT examination of the chest is similar to that found in contrast examination of the abdomen, involving induced visual neglect.

Keywords

Diagnostic radiology; observer performance; images; interpretation; quality assurance

INTRODUCTION

“What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need

© 2012 The Association of University Radiologists. Published by Elsevier Inc. All rights reserved.

Address correspondence to: Kevin Berbaum, PhD, Department of Radiology, 3170 Medical Laboratories, The University of Iowa, Iowa City, IA 52242, phone: 319-335-8122; fax: 319-356-2222, kevin-berbaum@uiowa.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

to allocate that attention efficiently among the overabundance of information sources that might consume it.”—Herbert A. Simon (1)

Satisfaction of search (SOS) error is a type of false negative error that occurs when multiple abnormalities are present on an examination and the radiologist fails to report at least one of them (2–7). SOS has been studied in numerous laboratory experiments (8–24) in which radiographs are read by radiologists, fellows or residents. ROC studies compared detection accuracy for *test* abnormalities that appeared on two types of radiographs: those with only the *test* abnormalities, and those with the *test* abnormalities and *added* abnormalities. All experiments demonstrated decreased reporting of *test* abnormalities when *added* abnormalities were present. Missed lesions are only attributed to SOS where it can be shown that they are detected when other lesions are absent. Laboratory studies of SOS have focused on three types of traditional imaging examinations: chest radiography (8–10,13,15–17,20,22), trauma radiography (12,18,21,23), and contrast studies of the abdomen (11,14,19). SOS effects have been studied using evidence from: receiver operating characteristic (ROC) curves (8–12,16,19–23), gaze-dwell time on missed abnormalities (13–15,17,18) and inspection time (9,10,19,23), and interventions to prevent SOS effects (10,19,20,22,24). A review summarizing research on SOS effects in radiographic modalities (25) is available.

The research findings suggest that there are two types of SOS effects. *Type I SOS effects*, found in chest and trauma radiology, appear to be the result of faulty pattern recognition (15,17,18). Decreases in ROC accuracy result from decreases in true-positive probability for each ROC point without changes in false-positive probability: the ROC points move straight down (8,10,12,16,21,23). This type of SOS is unrelated to changes in search behavior or inspection time except for the time involved in actually reporting abnormalities (9,10,13,15,17,18,21). *Type II SOS effects*, found in contrast studies of the abdomen, appear to be based on global changes in visual search behavior and faulty scanning. Reductions in true positive fractions are accompanied by reductions in false positive fractions: the ROC points move downward along the ROC curve (11,19). Although such movement is ordinarily indicative of reduced willingness to report abnormality, the cause of the reduced willingness to call abnormality in *Type II SOS* has been identified as reduced inspection of the non-contrast region with contrast studies (14).

Although we understand SOS effects in traditional radiographic imaging reasonably well (25), we know relatively little about SOS effects in advanced imaging (21,24). Clinical observations suggest a high frequency of SOS in CT imaging (26–29). For example, in a retrospective analysis of patients with cancers overlooked at CT, White (29) noted that 43% of patients had a major distracting finding. The goal in the current research was to determine the nature of SOS effects in CT of the chest.

EXPERIMENT 1

MATERIALS AND METHODS

Experiment 1 consisted of 70 CT examinations of the chest, half of which contained a native abnormality and half with no clinically significant abnormality present. These cases were presented to a group of readers made up of senior residents and fellows from the Department of Radiology. The readers saw each case twice in sessions separated by two months. Each case was seen with and without a pulmonary nodule present.

The experiment consisted of two parts separated by 2 months to reduce the likelihood that the study images and the responses to them would be remembered. Half of the cases presented in each part contained an *added* nodule and half did not. Thus, in the course of two parts, each CT examination appeared twice, once with and once without an *added* nodule.

Within each part, examinations were presented in a pseudorandom order so that occurrence of native *test* abnormalities and *added* nodules and normal radiographs would be unexpected.

Experimental Conditions—Laboratory studies 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. We compared the detection accuracy for various naturally occurring, subtle lesions on chest CT with accuracy for detecting those same lesions when simulated but realistic nodules were added (see Figure 1). The simulated nodules were designed to mimic primary lung neoplasms recognizing that in many instances, particularly small nodules, distinctions between malignant and benign lesions may be difficult or impossible. Thus, the background anatomy and native *test* lesions were perfectly matched for the two conditions compared in our experiment.

CT Examinations—CT examinations were obtained from clinical studies with approval by our local institutional review board. Verification of the lesions and disease state 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 multislice scanners using standard helical scanning parameters: 135 kVp, 250–260 mAs, 3-mm slice width, 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.

Native abnormalities include in this experiment were selected by chest radiologists (B.H.T., B.F.M., A.N.E.). They included: splenomegaly, thymic mass (2 examples), hiatal hernia (3 examples), thyromegaly, thyroid masses (4 examples), pleural thickening, adrenal mass, breast mass, pericardial effusion (3 examples), liver lesions (3 examples), and one each of biliary tree dilation, cirrhotic lesion, fatty liver, liver cyst, renal cyst, axillary node, absent spleen, retroperitoneal node, esophageal mass, left lower lobe air space disease (ground glass), renal calcification, right aortic arch, subcarinal nodes, aortic dilatation, and cardiophrenic node.

Examinations with no native abnormalities were age- and sex-matched to the examinations with native abnormalities. There were no other selection criteria. They contained normal variants that mimic disease no more frequently than would be expected in any sample of normal chest CT examinations from the clinic.

Image Display—Cases were presented on a CT workstation consisting of a Dell Precision 360 Mini-tower and two three-megapixel LCD monitors (National Display System) matched and calibrated. Monitors were calibrated to the DICOM standard using manufacturer's specifications. Periodically throughout the two-month course of the experiment the calibration was checked and when necessary adjusted.

A clinical display emulation software called WorkstationJ (30–32) (available from <http://perception.radiology.uiowa.edu/Software/WorkstationJ/tabid/236/Default.aspx>) was used to collect observer responses, including confidence ratings of abnormality and their locations, and display operations such as window and level adjustment and time on each axial CT slice. Visualization was timed for ranges of computed tomography (CT) slices with and without abnormalities.

Simulation of Pulmonary Nodules on CT Examinations of Patients

Lesion Removal: A thorough description of the procedures used to remove lesions and add nodules has been previously reported by Madsen *et al.*, (33). Briefly, software was developed using Interactive Data Language (IDL) to remove abnormalities such as nodules, granulomas, pleural tags, and cysts from CT images. The region occupied by the abnormality is replaced by normal tissue values sampled from the area surrounding the abnormality. The abnormality is removed by multiplying the complement of the lesion mask (i.e., $1 - \text{lesion mask}$) into the CT slice leaving a “hole” in the image. The background mask is also multiplied into the CT slice, which zeroes everything except the image values overlying the background mask. This result is shifted back to the abnormality coordinates and is added to the CT image filling in the void left by the first masking operation. Because the contours of the mask and its complement are gradual and exactly matched, the result contains no perceivable boundary effects.

Nodule Placement: In cases which already contained a pulmonary nodule, interactive software was developed to remove lesions and replace them with normal areas from the surrounding region (33). 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 and background regions. In cases without a pulmonary nodule present, software was developed that allowed an operator to add lesions from the database at any desired location and slice sequence of a CT series. Locations of placed nodules were selected by chest radiologists to be typical of nodules seen on clinical CT examinations.

Observers—Eighteen volunteer radiologists from our Department of Radiology were recruited as observers. They included seven fellows in post-residency training positions. Eleven second-, third- or fourth-year residents also agreed to participate. All subjects were given and signed an informed consent document that had been approved by our institutional review board for human subject use.

Procedure—Prior to the start of the experiment, each observer was read instructions and shown with a demonstration case how to display the images, make responses, and advance through the cases. Observers were instructed to search for all clinically significant abnormalities and to identify each abnormality by a mouse click over the abnormality. This produced a menu box for rating their confidence that the feature was abnormal using discrete terms such as “definitely abnormal,” “probably abnormal,” “possibly abnormal,” and “probably normal, but with some suspicion of abnormality.” They were also directed to rate their confidence that the feature was abnormal by using a subjective probability scale: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%. (No response to an abnormality was treated as “definitely normal” and 0% confidence. In our scoring, the report of abnormality with the highest confidence was considered false positive if it occurred on a case with no native abnormality.) Observers were instructed to first search the lung fields with window and level preset for lungs and afterward to continue searching using window and level presets for mediastinum and bone.

Data Analysis

ROC Analysis: We used the subjective probability scale which has 11 categories with both the empirical ROC method and the proper binormal model (PROPROC) (34) to study detection accuracy. For each treatment-reader combination, we computed area under the ROC curve (AUC) as our primary measure of detection accuracy. Experimental conditions were compared using an analysis of variance (ANOVA) (35, 36), to generalize to the population of radiologists (37).

Revision of Scoring Based on Review of CAD Findings: After this experiment was completed, we subjected our cases as seen in both experimental conditions to analysis by a commercial CAD system. Based on an extensive review of CAD findings, five of the 70 cases were judged to be compromised and removed from scoring. This included 3 cases with native abnormalities and 2 cases without native abnormalities. On these five cases, the examinations could have contained a nodule even in the control condition where nodules were supposed to be absent (the non-SOS condition) thereby compromising the experimental manipulation.

RESULTS

Empirical ROC areas for detection of native abnormalities were 0.55 without added nodules and 0.54 with the added nodules ($F(1,17) = 1.20, p = 0.2885$). Areas under the proper binormal ROC curves for detection of native abnormalities were 0.62 without added nodules and 0.60 with the added nodules ($F(1,17) = 1.99, p = 0.1764$).

DISCUSSION

An obvious factor limiting the magnitude of an SOS effect on detection accuracy was the difficulty of our *test* abnormalities. The average empirical ROC area was only 0.55 for detection of *test* abnormalities before the addition of nodules. There simply were not many examinations on which to measure SOS on accuracy because the *test* abnormality was missed even without the nodule being added. We also analyzed the rating data with other ROC models and accuracy parameters including partial ROC area and sensitivity at a fixed specificity. Conclusions were the same regardless of the ROC model or accuracy parameter.

The problem comes into clearer focus when we look at the ROC data matrices. Our best reader only found 16 of our 32 *test* abnormalities without the added nodule; two thirds of readers found less than half of that. Another way to look at the results is to count the number of readers (of 18) who found the *test* abnormality for each examination. Even without the nodule, about two-thirds of *test* abnormalities were missed by 80% of readers. The *test* abnormality was found in only 5 of 32 examinations by most readers.

In trying to account for this unexpected result, we questioned whether we had created such a strong expectation of the importance of detecting nodules that observers did not really look for other abnormalities. To test this possibility, we asked observers back for a supplementary experiment on the influence of clinical history. The delay between the end of the main experiment and the beginning of the follow-up experiment was about 8 months. Eleven of the original 18 observers agreed to participate. The 70 examinations, none of which contained nodules, were presented in two experimental conditions: once with and once without a clinical history directing search toward the native abnormality when it was present. We also removed the requirement to say whether the abnormality was a nodule when it was reported. There were two sessions with half of the examinations presented with histories and half without histories in each session. We assumed that history would improve detection of the *test* abnormalities, but also that detection of *test* abnormalities might improve even without the history. We expected that the presence of histories in the experiment and the lack of any nodules would reorient observers toward a more general search task. Indeed, in the control condition (without history and without the nodule), all readers found at least 8 of 32 *test* abnormalities. History did not seem to help readers much further: in the experimental condition aided by a suggestive history, the best reader now found 17 of 32 *test* abnormalities. Viewed another way, without history, most *test* abnormalities were missed by most readers; but with history, about a third of *test* abnormalities were found by most readers. The modest improvement in performance with

history suggests that the low detection of *test* abnormalities in our main experiment was probably not based on readers looking only for nodules.

This result was unexpected. Two senior chest radiologists each with more than ten years of faculty experience selected roughly 400 patient examinations from examinations interpreted in our chest section over the course of three years that were considered for inclusion based on the presence of subtle findings and the lack of multiple disease processes. These examinations were further reviewed by two senior residents working with the two senior radiologists to select examinations with abnormalities that would be expected to be appreciated under ordinary conditions. A subtle abnormality was defined as one that might not always be found.

Perhaps our reader sample should have included more faculty radiologists and excluded second-year residents. The second-year residents were only allowed to participate late in their second year, having had at least two chest rotations with the second focusing heavily on CT. In our program, third-year residents have an additional rotation focusing on CT which completes their training in the interpretation of chest CT. We could have included several faculty radiologists expert in chest CT interpretation but naïve to our patient examinations. However, we have discovered that doing so would not have improved the low detection rate very much. After the first experiment was completed, we asked a senior chest radiologist to review the examinations that had appeared without nodules. Under blinded conditions, this advanced reader found only 10 of the 32 test abnormalities. Thus our faculty chest radiologist performed about the same or was outperformed by 6 of our Experiment 1 readers. This finding convinces us that the test abnormalities were too difficult even for chest CT faculty.

On reviewing the unreported abnormalities, we found that some, although representing true radiologic abnormalities, may not have had a significant bearing on the patient's health status. In seeking subtle abnormalities, we may have strayed into including some non-significant *test* abnormalities.

The '20/20 hindsight' or 'retrospectroscope' problem may be worse in CT examinations because lesions are more visible in CT slices than in radiographs. CT avoids the superposition of structures inherent in radiography, allowing smaller abnormalities to be appreciated, at the cost of extensive navigation. Exhaustive search may not be possible in a practical time frame. It is easy to forget the length of the slice-by-slice navigation though the volume required to visualize the lesion. As such, navigation time needed to arrive at the slice containing abnormality may indicate lesion subtlety on CT imaging.

Because of the possibility of a "floor" effect – not enough examinations on which to measure SOS detection accuracy because the *test* abnormality was missed even without the nodule being added – the experiment needed to be repeated with more detectable and important abnormalities as *test* abnormalities.

EXPERIMENT 2

MATERIALS AND METHODS

The experimental conditions, procedure, CT and display equipment, and methods for simulation of nodules of Experiment 2 were the same as for Experiment 1. However, new cases were included with more obvious *test* abnormalities, and presented to a new sample of 14 readers made up of senior residents and fellows from the Department of Radiology. Different readers for Experiment 2 were available because it was conducted two and a half years after the first experiment.

Experiment 2 consisted of 68 CT examinations of the chest, 33 of which contained a native abnormality and 35 with no clinically significant abnormality present. (All compromised examinations in the first experiment were eliminated.) Readers saw each case twice in sessions separated by two months. Each case was seen with and without a pulmonary nodule present.

CT Examinations—Again, CT examinations were obtained from clinical studies with approval by our local institutional review board. Verification of the lesions and disease state was by the same method as Experiment 1, and patient identifiers were removed.

All 35 CT examinations without native abnormalities used in Experiment 1 were also used in Experiment 2. Of the 35 CT examinations with native abnormalities used in Experiment 1, 12 with the highest rates of detection were retained for Experiment 2. These consisted of liver lesion (2 examples), adrenal mass, breast lesion, hiatal hernia, absent spleen, thyroid lesion (2 examples), cirrhotic liver, renal calcification, aortic arch anomaly, and subcarinal nodes. Twenty-one new CT examinations with native abnormalities included: thymic node, enlarged ascending aorta, aortic aneurysm, cardiomegaly, gallstone (2 examples), pleural plaques, adrenal myelolipoma, adrenal mass (3 examples), gastric mass, precarinal node, hiatal hernia, hiatal hernia with thickened esophageal tissue, mediastinal mass, enlarged pulmonary artery, abdominal ventral hernia, pulmonary embolus (2 examples), and kidney cyst.

Observers—Fourteen volunteer radiologists included two fellows in post residency training positions and 12 second-, third- or fourth-year residents. All subjects were given and signed an informed consent document that had been approved by our institutional review board for human subject use.

Procedure—The only difference in procedure between Experiments 1 and 2 was that in Experiment 1, readers were asked to identify whether an abnormality reported was a pulmonary nodule or not using a radio button in the menu but in Experiment 2 readers were asked to type a brief description of the reported abnormality in the text box.

Data Analysis

ROC Areas: We used the subjective probability scale to study ROC area and the natural language categorical scale to study decision thresholds. With 11 categories, the subjective probability scale may offer more operating points (38), whereas the natural language based scale may offer more stable operating points (11,14,19). A small number of discrete categories defined using natural language terms provides the best chance of category boundaries to have fixed boundaries in the observer's memory (19). Analysis of ROC area was similar to that already reported for Experiment 1 except that we used analysis of variance with a between-subject factor for experiment and a within-subject factor for experimental condition to study detection accuracy across the two experiments (35, 36).

Decision Thresholds: To test for a *Type II SOS effect* on decision thresholds, we used the five-point natural language scale. One of the best and simplest indices of decision thresholds is the false-positive fraction (FPF) associated with each ROC point (39). In our scoring, a report of abnormality was considered a false positive if it occurred on a case with no native abnormality. False positive fractions associated with ROC points were studied using an analysis of variance with a between-subject factor for experiment and within-subject factors for experimental condition (non-SOS, SOS), and threshold (achieved by grouping the 5-point ratings in 4 ways: 1234/5, 123/45, 12/345, 1/2345). True positive fractions associated

with ROC points were studied using a similar analysis of variance with a between-subject factor for experiment and within-subject factors for experimental condition, and threshold.

Inspection Time: As already noted, in *Type II SOS*, there is a shift in decision thresholds toward greater conservatism (11,19) that has been linked to reduced inspection of examination areas outside the region of the abnormalities (14). If decision thresholds shift in the current experiment, we need to test whether there is also reduced inspection on examination regions that contain no abnormalities. WorkstationJ measured observer inspection time on each CT slice. We analyzed the visualization time within slice ranges containing (1) the *test* native abnormalities, (2) the added nodules, and (3) no abnormalities. Because different examinations were used in the first and second experiments, we computed median inspection time across examinations without native abnormalities and across examinations with native abnormalities for each reader-treatment combination. Reader medians for each of the three CT slice ranges were analyzed with an ANOVA with a within-subject factor for treatment and a between-subject factor for experiment.

RESULTS

Detecting Native Abnormalities

Detection Accuracy: For Experiment 2, empirical ROC areas for detection of native abnormalities were 0.63 without added nodules and 0.64 with the added nodules ($F(1,13) = 0.20$, $p = 0.6637$). For Experiment 2, areas under the proper binormal ROC curves for detection of native abnormalities were 0.68 without added nodules and 0.71 with the added nodules ($F(1,13) = 1.35$, $p = 0.2668$). An analysis of variance that examined empirical ROC area from both experiments resulted in a difference in detection accuracy between the experiments (Experiment 1 = 0.54 vs. Experiment 2 = 0.64, $F(1,30) = 46.81$, $p < 0.0001$), but no significant difference for experimental condition and no significant interaction. An analysis of variance that examined proper binormal model ROC area from both experiments resulted in a difference in detection accuracy between the experiments (Experiment 1 = 0.61 vs. Experiment 2 = 0.70, $F(1,30) = 14.78$, $p = 0.0006$), but no significant difference for condition and no significant interaction.

Decision Thresholds: Figure 2 presents plot of ROC points from the five-point subjective probability scale for reporting *test* abnormalities from the two experiments. The points suggest that Experiment 2 achieved more accurate detection than did Experiment 1 and a great willingness to call abnormality in Experiment 2. Both experiments show more conservative decision thresholds in the SOS condition than the non-SOS condition with fewer true-positives and fewer false positives.

The analysis of variance of false-positive fractions associated with ROC points yielded statistically significant differences between the first and second experiments (0.17 vs. 0.26, $F(1,30) = 4.67$, $p = 0.0389$) and between non-SOS and SOS conditions (0.23 vs. 0.20, $F(1,30) = 5.49$, $p = 0.0260$). Decision thresholds were less conservative in Experiment 2 than Experiment 1, but more conservative in the SOS condition than the non-SOS condition.

The analysis of variance of true-positive fractions associated with ROC points yielded statistically significant differences between the first and second experiments (0.25 vs. 0.50, $F(1,30) = 38.16$, $p < 0.0001$) and between non-SOS and SOS conditions (0.39 vs. 0.36, $F(1,30) = 4.26$, $p = 0.0478$). In general, true positive fractions for finding native abnormalities were higher in the second experiment than the first experiment, but were lower in the SOS condition than the non-SOS condition. Thus, changes in false positive fractions were mirrored in true-positive fraction consistent with decision threshold shifts.

Inspection Time: Image visualization times were compared for two situations—with and without the added nodule. We focused on examinations without native abnormalities in order to understand difference in visual search underlying changes in decision thresholds. For the CT slice range that contained the nodule (when it was present), the average inspection time was 3.6 seconds without nodules and 12.7 seconds with nodules. For CT slices that contained no abnormalities, the average inspection time was 124.0 seconds without nodules and 112.0 seconds with nodules. Thus, the inspection time for the nodule region increased by 9 seconds with the addition of the nodule ($F(1,30) = 209.16$, $p = 0.0001$), but inspection time for the region without abnormalities decreased by 12 seconds ($F(1,30) = 13.13$, $p = 0.0011$). Total inspection time for all slices was not different for the two conditions (128.1 seconds without nodules vs. 124.1 seconds with nodules, $F(1,30) = 1.60$, $p = 0.2160$). Thus, although overall inspection time is constant, increased inspection time on the nodule region when nodules are present leads to reduced inspection time elsewhere. There was a 10% reduction in visual search for the slice range without abnormalities from which false positives responses are counted for the ROC analysis. A similar pattern of results were found for examinations with native abnormalities.

DISCUSSION

A critical issue facing radiology is growth both in the volume of imaging studies being performed (40, 41) and the increasing volume of data collected in each study (42–44). There is so much image data that an exhaustive inspection of it may not be possible in any practical time frame. With so many images and ways of translating data into actual images (e.g., use of various reconstruction algorithms) in a single study, human attention becomes the limiting factor in diagnostic performance. We know relatively little about how radiologists cope with information load in advanced imaging or the blind spots in attention that may be created by these coping strategies.

The increased ROC area in the second experiment shows that we succeeded in providing more detectable *test* abnormalities in that experiment. There was no significant difference in ROC area between non-SOS and SOS conditions for either experiment or both experiments taken together. Decreases in both true- and false-positive fractions canceled each other in terms of effect on ROC area, but indicate shifts in decision thresholds with greater conservatism in the SOS condition. This conservatism was accompanied by a global change in visual search of the chest CT examinations with reduced search time on CT slice ranges without abnormalities in examinations without native abnormalities. The reduction was about 10% of search time in CT slice ranges without abnormality.

This pattern of results is characteristic of *Type II SOS* noted in contrast radiography of the abdomen (11, 14, 19). Classic studies found *Type II SOS effects* on the report of abdominal lesions visible without contrast: readers more often missed plain film abnormalities present on contrast studies but also made fewer false-positive (FP) responses (11,19). In gaze-dwell time studies, a global change in search behavior for the contrast studies was found — a visual neglect of the non-contrast regions that could explain the combined reduction in true- and false-positive responses (14). When contrast was introduced, readers focused on the contrast and neglected plain film regions leading to reductions in both true- and false-positive response to plain film regions. Despite the similarities between the current findings and those of classic *Type II SOS* studies, there is an important difference. Whereas the classic *Type II SOS effect* – a visual neglect of plain film regions – is often nearly complete, the visual neglect noted in the current research is limited in magnitude by the amount of search time taken to process the added abnormality. The inspection time appears a constant with time taken by an added abnormality subtracted from the time remaining to process the remainder of the examination.

One further point must be taken into account in interpreting our results: total inspection time per case may have been much less than in the clinical situation. Average inspection time was just over 2 minutes per examination when no native abnormalities were present and just under 2 and a half minutes per examination when native abnormalities were present. Residents working in our chest CT reading area report taking about 5 minutes per examination on average, but taking as much as 10 or 15 minutes for a complicated case. One difference is that our reporting method is much more time efficient: the reader only has to click on the abnormality and work through a pop-up menu of a few radio buttons. At most, only typing a short phrase describing the abnormality is required. In the clinic, a radiology report must be dictated and corrected. This may account for some of the difference. Another way to think about laboratory vs. clinical reading is to consider the case load. Our experiments each contained about 70 CT examinations in each half of each experiment. We did not require all examination to be read in a single setting, several sessions were sometimes required. Each half took perhaps 3 hours to complete. On the other hand, daily case load for a faculty radiologist in the chest room might be about 150 radiographs and 25 CTs. They would typically take less than a minute of actual reading time sans reporting for a radiograph and about 2–3 minutes for a CT, longer to make comparisons with older studies. Third- and fourth-year residents and fellows typically take twice as long to read chest CT examinations with an output of perhaps 12 CT cases in a day rather than the 25 that the faculty may do. Our case sample is as large as it is because of the demands of ROC methodology. The 70 cases needed for ROC method, read in three or four hours for each half of one of our experiments, would be more CTs than a senior resident or fellow would encounter in a week in our chest room.

The readers were not rushed or limited in reading time in any formal way. In each scheduled session, they were able to read however many cases for however long they wanted to stay. They could also schedule as many sessions as they needed. Of course, they were aware that they were participating in a laboratory study, that their judgments would have no effect on patient care and that their individual performance would not be disclosed to anyone. They were also paid for their participation at a fixed rate, so the longer they took to read the examinations, the less they were paid for their time.

We do not know whether the SOS effects observed in the current experiments depends on our high caseload. To find out, we would need to repeat our study with similar readers taking the better part of a week to read each half of the study. At present, this does not fit within the residency schedule. On the other hand, our experiments probably do generalize to the kind of high workload situation that is increasing of concern in radiology (40–46).

Acknowledgments

Supported by USPHS Grant R01 EB 00145 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), Bethesda, Maryland.

References

1. Simon, HA. Designing organizations for an information-rich world. In: Greenberger, Martin, editor. *Computers, Communications, and the Public Interest*. Johns Hopkins Press; Baltimore: 1971. p. 37-52.
2. Tuddenham WJ, Calvert WP. Visual search patterns in roentgen diagnosis. *Radiology*. 1961; 76:255–256. [PubMed: 13778547]
3. Tuddenham WJ. Problems of perception in chest roentgenology: Facts and fallacies. *Radiol Clin North Am*. 1963; 1:227–289.

4. Tuddenham WJ. Visual search, image organization and reader error in roentgen diagnosis. *Studies of the psychophysiology of roentgen image perception*. Radiology. 1962; 78:694–704. [PubMed: 13923013]
5. Smith, MJ. *Error and Variation in Diagnostic Radiology*. Springfield, IL: Charles C. Thomas; 1967. p. 27
6. Renfrew DL, Franken EA Jr, Berbaum KS, Weigelt FH, Abu-Yousef MM. Error in radiology: Classification and lessons of 182 cases presented at a problem case conference. *Radiology*. 1992; 183:145–150. [PubMed: 1549661]
7. Ashman CJ, Yu JS, Wolfman D. Satisfaction of search in osteoradiology. *Am J Roentg AJR*. 2000; 175:541–544.
8. Berbaum KS, Franken EA Jr, Dorfman DD, et al. Satisfaction of search in diagnostic radiology. *Invest Radiol*. 1990; 25:133–140. [PubMed: 2312249]
9. Berbaum KS, Franken EA, Dorfman DD, et al. Time-course of satisfaction of search. *Invest Radiol*. 1991; 26:640–648. [PubMed: 1885270]
10. Berbaum KS, Franken EA Jr, Anderson KL, et al. The influence of clinical history on visual search with single and multiple abnormalities. *Invest Radiol*. 1993; 28:191–201. [PubMed: 8486483]
11. Franken EA Jr, Berbaum KS, Lu CH, et al. Satisfaction of search in detection of plain film abnormalities in abdominal contrast examinations. *Invest Radiol*. 1994; 29:403–409. [PubMed: 8034444]
12. Berbaum KS, El-Khoury GY, Franken EA Jr, et al. Missed fractures resulting from satisfaction of search effect. *Emerg Radiol*. 1994; 1:242–249.
13. 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. [PubMed: 7862998]
14. Berbaum KS, Franken EA Jr, Dorfman DD, et al. Cause of satisfaction of search effects in contrast studies of the abdomen. *Acad Radiol*. 1996; 3:815–826. [PubMed: 8923900]
15. Berbaum KS, Franken EA Jr, Dorfman DD, et al. The role of faulty visual search in the satisfaction of search effect in chest radiology. *Acad Radiol*. 1998; 5:9–19. [PubMed: 9442202]
16. Berbaum KS, Dorfman DD, Franken EA Jr, Caldwell RT. Proper ROC analysis and joint ROC analysis of the satisfaction of search effect in chest radiography. *Acad Radiol*. 2000; 7:945–958. [PubMed: 11089697]
17. Berbaum KS, Franken EA Jr, Dorfman DD, et al. The role of faulty decision making in the satisfaction of search effect in chest radiography. *Acad Radiol*. 2000; 7:1098–1106. [PubMed: 11131054]
18. Berbaum KS, Brandser EA, Franken EA Jr, et al. Gaze dwell times on acute trauma injuries missed because of satisfaction of search. *Acad Radiol*. 2001; 8:304–314. [PubMed: 11293778]
19. Berbaum KS, Franken EA Jr, Dorfman DD, et al. Can order of report prevent satisfaction of search in abdominal contrast studies? *Acad Radiol*. 2005; 12:74–84. [PubMed: 15691728]
20. Berbaum KS, Franken EA, Caldwell RT, Scharzt KM. Can a checklist reduce SOS errors in chest radiography? *Acad Radiol*. 2006; 13:296–304. [PubMed: 16488841]
21. Berbaum KS, El-Khoury GY, Ohashi K, Scharzt KM, Caldwell RT, Madsen M, Franken EA Jr. Satisfaction of search in multitrauma patients: Severity of detected fractures. *Acad Radiol*. 2007; 14:711–22. [PubMed: 17502261]
22. Berbaum KS, Caldwell RT, Scharzt KM, Thompson BH, Franken EA Jr. Does computer-aided diagnosis for lung tumors change satisfaction of search in chest radiography? *Acad Radiol*. 2007; 14:1069–76. [PubMed: 17707314]
23. Berbaum KS, Scharzt KM, Caldwell RT, El-Khoury GY, Ohashi K, Madsen M, Franken EA. Satisfaction-of-search for subtle skeletal fractures may not be induced by more serious injury. *JACR*. 2012; 9:344–351. DOI: 0.1016/j.jacr.2011.12.040. [PubMed: 22554633]
24. Scharzt K, Berbaum K, Madsen M, Thompson B, Mullan B, Caldwell R, Hammett B, Ellingson A, Franken E. Multiple diagnostic task performance in computed tomography examination of the chest. *BJR*. 2012 in press.
25. Berbaum, KS.; Franken, EA., Jr; Caldwell, RT.; Scharzt, KM. Satisfaction of search in traditional radiographic imaging. In: Samei, E.; Krupinski, E.; Jacobson, F., editors. *The Handbook of*

- Medical Image Perception and Techniques. Cambridge, UK: Cambridge University Press; 2010. p. 107-138.
26. Gurney JW. Missed lung cancer at CT: Imaging findings in nine patients. *Radiology*. 1996; 199:117–122. [PubMed: 8633132]
 27. Kakinuma R, Ohmatsu H, Kaneko M, Eguchi K, Naruke T, Nagai K, Nishiwaki Y, Suzuki A, Moriyama N. Detection failures in spiral CT screening for lung cancer: Analysis of CT findings. *Radiology*. 1999; 212:61–66. [PubMed: 10405721]
 28. Davis SD. Through the “retrospectroscope”: A glimpse of missed lung cancer in CT. *Radiology*. 1996; 199:23–24. [PubMed: 8633150]
 29. White CS, Romney BM, Mason AC, et al. Primary carcinoma of the lung overlooked at CT: Analysis of findings in 14 patients. *Radiology*. 1996; 199:109–115. [PubMed: 8633131]
 30. Scharztz, KM.; Berbaum, KS.; Caldwell, RT.; Madsen, MT. WorkstationJ: workstation emulation software for medical image perception and technology evaluation research. In: Jiang, Yulei; Sahiner, Berkman, editors. *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment*; Proceedings of SPIE Vol. 6515; Bellingham, WA: SPIE; 2007. p. 65151I
 31. Scharztz, KM.; Berbaum, KS.; Caldwell, RT.; Madsen, MT. WorkstationJ as ImageJ plugin for medical image studies. Annual Meeting of the Society for Imaging Informatics in Medicine (SIIM) – 9th Annual SIIM Research & Development Symposium; Charlotte, NC. June 6, 2009; (<http://www.siimweb.org/assets/FCBE219A-C30B-4003-9892-FACA9230AB91.pdf>)
 32. Krupinski EA, Berbaum KS, Caldwell RT, Scharztz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. *J Am Coll Radiol*. 2010; 7:698–704. [PubMed: 20816631]
 33. 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; 13:305–312. [PubMed: 16488842]
 34. Pesce LL, Metz CE. Reliable and computationally efficient maximum-likelihood estimation of “proper” binormal ROC curves. *Acad Radiol*. 2007; 14:814–829.10.1016/j.acra.2007.03.012 [PubMed: 17574132]
 35. Kirk, RE. *Experimental Design*. 2. Vol. 75. Belmont, CA: Wadsworth; 1982. p. 429-455.p. 531-548.
 36. Dixon, WJ. *Accompanying BMDP Release 7*. Vol. 2. Berkeley: University of California Press; 1992. BMDP Statistical Software Manuel; p. 1239-1258.p. 1431-1433.
 37. Berbaum KS, Dorfman DD, Franken EA. Measuring observer performance by ROC analysis: Indications and complications. *Invest Radiol*. 1989; 24:228–233. [PubMed: 2753639]
 38. Berbaum KS, Dorfman DD, Franken EA Jr, Caldwell RT. An empirical comparison of discrete ratings and subjective probability ratings. *Acad Radiol*. 2002; 9:756–63. [PubMed: 12139089]
 39. Swets, JA.; Pickett, RM. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press; 1982. p. 39
 40. Hillman BJ. Economic, legal, and ethical rationales for the ACRIN National Lung Screening Trial of CT screening for lung cancer. *Acad Radiol*. 2003; 10:349–350. [PubMed: 12643562]
 41. Rothenberg BM, Korn A. The opportunities and challenges posed by the rapid growth of diagnostic imaging. *JACR*. 2005; 2:407–410. [PubMed: 17411844]
 42. Andriole KP, Morin RL, Arenson RL, Carrino JA, Erickson BJ, Horii SC, Piraino DW, Reiner BI, Seibert JA, Siegel E. Addressing the coming radiology crisis—The Society for Computing Applications in Radiology Transforming the Radiological Interpretation Process (TRIP). Initiative J Digit Imaging. 2004; 17:235–243.
 43. Andriole KP, Morin RL. Transforming medical imaging: The first SCAR TRIPTM conference: A position paper form the SCAR TRIPTM Subcommittee of the SCAR Research and Development Committee. *J Digit Imaging*. 2006; 19:6–16. [PubMed: 16511675]
 44. Andriole KP, Wolfe JM, Khorasani R, Treves ST, Getty DJ, Jacobson FL, Steigner ML, Pan JJ, Sitek A, Seltzer SE. Optimizing analysis, visualization, and navigation of large image data sets: One 5000-section CT scan can ruin your whole day. *Radiology*. 2011; 259:346–362. [PubMed: 21502391]

Figure 1.

The abnormalities used in case 56 shown using a lung window/level in the left panel and a mediastinal window/level on the right panel. The left panel shows a pulmonary nodule (indicated by the white box) which was present in slices 121–127 of the CT examination in the SOS experimental condition, but was not present in the CT examination in the non-SOS control condition. The right panel shows the *test* abnormality (indicated by the white box), an example of aortic dilatation, which was visible in slices 67–99 of the CT examination in both experimental conditions.

Figure 2.

Plot of ROC points for detecting *test* abnormalities, diverse subtle abnormalities, from two experiments. Experiment 1 points are triangles with black triangles representing the non-SOS condition and white triangles representing the SOS condition. Experiment 2 points are circles with black circles representing the non-SOS condition and white circles representing the SOS condition. Clearly Experiment 2 achieved more accurate detection than did Experiment 1. Both experiments show a shift in decision thresholds toward greater strictness in the SOS condition with fewer true-positives and fewer false positives. This decision threshold shift is indicative of less visual search.