Satisfaction of Search in Multi-trauma Patients: Severity of Detected Fractures

Kevin S. Berbaum, PhD, **George Y. El-Khoury, MD**, **Kenjirou Ohashi, MD**, **Kevin M. Schartz, PhD**, **Robert T. Caldwell, MFA**, **Mark Madsen, PhD**, and **Edmund A. Franken Jr., MD**
1From the Department of Radiology, The University of Iowa Roy J. and Lucille A. Carver College of Medicine, Iowa City, IA 52242.

Abstract

Purpose—Satisfaction of search (SOS) occurs when an abnormality is missed because another abnormality has been detected. This research studied whether the severity of a detected fracture determines whether subsequent fractures are overlooked.

Materials and Methods—Each of seventy simulated multi-trauma patients presented examinations of three anatomic areas. Readers evaluated each patient under two experimental conditions: when the images of the first anatomic area included a fracture (the SOS condition), and when it did not (the control condition). The SOS effect was measured on detection accuracy for subtle test fractures presented on examinations of the second and third anatomic areas. In an experiment with twelve radiology readers, the initial SOS radiographs showed non-displaced fractures of extremities, fractures associated with low morbidity. In another experiment with twelve different radiology readers, the initial examination, usually a CT, showed cervical and pelvic fractures of the type associated with high morbidity. Because of their more direct role in patient care, the experiment using high morbidity SOS fractures was repeated with seventeen orthopedic readers.

Results—Detection of subtle test fractures was substantially reduced when fractures of low morbidity were added (*p*<0.01). No similar SOS effect was observed in either experiment in which added fractures were associated with high morbidity.

Conclusion—The satisfaction of search effect in skeletal radiology was replicated, essentially doubling the evidence for SOS in musculoskeletal radiology, and providing an essential contrast to the absence of SOS from high morbidity fractures.

INTRODUCTION

Physicians have long been aware that an injury may draw and hold their attention, diverting it from other injuries (1). A "satisfaction of search" (SOS) effect has been demonstrated in which the discovery of a fracture on one image interfered with the detection of a subtle fracture on another image of the same patient (2). Detection of subtle *test* fractures was substantially reduced when *additional* fractures were included in other images of each multi-trauma series. The clinical practice of radiology has changed greatly since this experiment was performed in 1994. Whereas the previous experiment used radiographs interpreted at a film viewer, current practice relies heavily on computed tomography and direct digital radiography with

interpretation at workstations equipped with high-resolution displays. The first experiment reported here attempts to replicate the SOS effect of finding "minor" *added* fractures on subsequent *test* fractures in a patient's multi-trauma series using modern images and displays.

In 2001, gaze time on fractures was measured in an attempt to determine whether *test* fractures were missed in this SOS effect because of misdirection of attention (3). Although readers spent somewhat less time inspecting subsequent radiographs of a patient's trauma series after viewing initial radiographs that contained added fractures, they generally did look at the test fractures that they failed to report. When the experiment was repeated to examine whether the severity of the added fracture affected search, *test* fractures were missed more often when they appeared with major *added* fractures than with minor *added* fractures. Because there were only ten cases and only one that included no test fracture, true- and false-positive rates could not be estimated with acceptable precision or certainty. We address this question in the second and third experiments using receiver operating characteristic (ROC) methodology, testing whether *added* fractures with major morbidity yield greater SOS.

In our second experiment, we changed the significance of the *added* abnormalities without disturbing the *test* abnormalities, allowing us to test whether detection accuracy for *test* fractures depends on the clinical significance of *added* fractures. Minor *added* fractures of our first experiment were non-displaced fractures of the hands, wrists, feet, ankles, ribs, shoulder, etc. that were detected on ordinary digital radiographs. Major *added* fractures of our second experiment involved potential major morbidity (such as a cervical spine fracture), usually involving computed tomography (CT). Multidetector CT has revolutionized the diagnostic evaluation of trauma and emergency room patients (6–10), particularly for spinal injury. Patients with life threatening spinal or pelvic injuries are usually examined using multidetector CT rather than with radiography. We expected that SOS effects on ROC area would be greater for major *added* fractures than minor *added* fractures. Readers of both the first and second experiments consisted of radiology residents and fellows.

It has long been suspected that clinical necessity underlies diagnostic oversight in the multiply injured patient (4,5). Fractures having immediate implications for patient care would not only have stronger SOS effects than minor fractures, but would have a greater impact on those physicians principally responsible for patient care. Surgeons or emergency room physicians may be more affected by SOS where discovery of a serious injury requires immediate treatment. In our third experiment, orthopedic surgery residents and fellows read the same case set as the radiology residents and fellows of our second experiment.

MATERIALS AND METHODS

We performed three ROC experiments that tested SOS effects in skeletal trauma radiology, each of which involved two experimental conditions. For each patient, the first examination displayed was normal for one experimental condition (the control condition) but included an *added* fracture in the other experimental condition (the SOS condition). Addition of fractures into the first examination was an experimental manipulation; we measured detection of the *test* fractures appearing in the second and third examinations, and gathered false-positive responses when both the second and/or third examinations were normal.

Imaging Material

Case material for this study came from existing radiology records. All patient identifiers were removed except for age, gender, and diagnosis. Our institutional review board granted exclusion from human use with regard to permission to use these records consistent with exemption 4 of National Institutes of Health human use policy.

To simulate the radiologic examination of the multi-trauma patient, each patient in the experiments consisted of imaging studies of three different body parts. Seventy simulated multi-trauma patients were constructed. Because our experiments involve only SOS and not broader aspects of imaging diagnosis, our case mix did not include extraskeletal abnormalities. Although the examinations of each patient came from different sources, they were matched so that they would appear to belong to the same patient. Raw material came from digital radiographs from 304 actual patients presenting over 800 normal or abnormal examinations. Additionally, over 200 patients presented normal or abnormal CT examinations of either the spine or pelvis. To the extent possible, we used examinations from the same patient. Where this was not possible, we matched examinations by gender and age. Each simulated patient was reviewed by our chief musculoskeletal radiologist who judged whether the examinations appeared to have come from the same patient. When the exams did not match well enough, we revised the composition of the simulated patient until that radiologist was convinced that another musculoskeletal radiologist could not detect that examinations from multiple patients were included in the series of examinations we presented as coming from a single patient. Next, a different musculoskeletal radiologist confirmed that the examinations presented as a patient gave no hint of coming from multiple patients. We made further revisions whenever this requirement was not met.

The case material for our experiments was derived from clinical imaging data in DICOM format. Software was developed to discard the DICOM header and convert DICOM computed tomography (CT) studies into volumetric, binary files. Each CT study thus became one large file consisting of N consecutive 512×512 16 bit images, where N is the number of frames in the CT series. Image parameters, such as pixel size, slice separation and the number of frames, were stored in a single ASCII descriptor file. Software was developed to generate sagittal and coronal reconstruction images with the correct aspect ratio based on the image parameters. Other software was developed to generate target image format files (TIFF) from DICOM format digital radiographs and to optimally resize them to fill the display screen. The dynamic range of the digital radiographs was adjusted to match the dynamic range of the CT studies so that the range went from a pixel value of 0 to 4095.

Simulated Multi-trauma Patients

Figures 1, 2, and 3 illustrate a single simulated patient from our experiment, case 28. Figure 1 provides enlargements demonstrating the fractures used in simulated case 28 to create various experimental conditions. Figure 2 shows the thumbnails in the initial display of case 28 as it appeared in different experiments and experimental conditions. Figure 3 shows how case 28 appeared with a major fracture present on the two-monitor display to a reader working through the examinations of the series.

Multi-trauma patients were depicted in a series of three examinations, each of a separate body part. When displayed (see below), the three examinations were presented in a specific order in which the *added* fracture could only appear in the first examination and the *test* fracture could appear in either the second or third examination of the series. This display procedure insured that images with *added* fractures would appear before those with *test* fractures. In the previous study (2), *test* and *added* fractures were randomly distributed in the simulated series and so could appear in either order. Assuming that discovery of the *added* fracture causes subsequently viewed *test* fractures to be overlooked; the current procedure of presenting *added* fractures first affords the chance that each simulated patient may contribute to an observed SOS effect. In clinical practice, a cervical spine CT examination ordinarily would be reviewed before other examinations because of the seriousness of a potential cervical spine injury.

Readers interpreted each series under two experimental conditions: when the first examination in the series included a fracture (the SOS condition), and when it did not (the non-SOS or

control condition). In the first experiment, the first examination in the series for each patient showed a minor non-displaced fracture as shown in detail in Figure 1C. Figure 2A shows how that examination appeared in the patient series to readers when it was presented to them as thumbnail images. In the second and third experiments, the first examination in the series for each patient showed fractures with major morbidity as shown in detail in Figures 1A and B. Figure 2C shows how that examination appeared in the patient series to readers when it was presented to them as thumbnail images. For the non-SOS condition of each experiment, a normal control exam of the same area, using the same technology was substituted for the initial examination as shown in Figure 2B and D.

In our first experiment, the imaging examination was designed to simulate the multi-trauma patient whose life was not threatened by injuries. We presented radiographs examining three different body parts, excluding the spine. The *added* fracture was always presented as the first examination in the series. In our second and third experiments, the imaging examination was designed to simulate multi-trauma patients whose life might be threatened by their injuries. So for the initially presented examination in the series, an examination of the spine or pelvis, usually a CT (as in Figure 2C and D), was substituted for the radiographic examination used in the first experiment (as in Figure 2A and B). In changing the significance of the added fractures, we had to change the anatomy being examined. For example, a pelvic radiograph in Figure 2A can contain a minor fracture of the acetabular bone, but a cervical spine study in Figure 2C is needed to present a fracture that can produce quadriplegia. Different examinations were needed for the control condition without added fractures (Figure 2B and D). The examination with the added fracture and the normal examination that it replaced were of the same anatomic region (as in the pelvis of Figures 2A and B, and the cervical spine of Figures 2C and 2D). The initial images in each simulated patient in our second and third experiments usually consisted of a multidetector CT of the cervical spine or pelvis and radiographs of two different body parts. In occasional cases, however, radiographs alone were used to examine the spine.

Detection accuracy was measured by scoring responses on the second and third examinations (e.g. the foot and chest radiographs in Figure 2A–D). The second and third body parts presented for each simulated patient were digital radiographs of extremities, chest, or pelvis and usually included multiple views. The same examinations of the second and third body parts were always presented for each patient in both control and experimental conditions of all three experiments. There were 27 simulated patients in which both the second and third examinations contained only normal digital radiographs, and 43 with a subtle test fracture. For each simulated patient, all responses to the second and third examinations were counted toward the patient rather than toward the examination so that there were 70 scored patients with 27 normals and 43 abnormals for purposes of ROC analysis.

Display

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

Software was developed from *ImageJ*, a public domain image processing and analysis package written in the Java programming language (available at ***http://rsb.info.nih.gov/ij***). This was accomplished by modifying and extending *ImageJ* so that it had capabilities to display both single radiographic images and voxel image stacks generated by computed tomography. This software also collected reader responses, not only their explicit judgments indicating detection and localization of abnormality and associated confidence rating, but also the display functions that they used to explore the image data including window and level adjustment, navigation

and time each image or CT slice was displayed. A log of the display operations and the resulting displays as a function of time was recorded.

Each patient's imaging examinations were first presented on the left monitor of a dual-monitor display station (Figure 3A). The reader saw the patient's age and gender, the case number, and thumbnail images of the studies comprising the case. (Figure 3A is presenting the thumbnails of the examinations shown in Figure 2C.) Readers were instructed that these images were not meant to be diagnostic, but to give an impression as to the number and type of examinations available for the patient. The actual diagnostic reading was done on the right monitor (Figure 3B–D). A small menu at the bottom of the right monitor permitted the reader to display the images, one examination at a time, on the right monitor using "next image" and "previous image" buttons. With these buttons, the reader could maximize the first examination on the right monitor as shown in Figure 3B, then maximize the second examination on the right monitor as shown in Figure 3C, and then maximize the third examination on the right monitor as shown in Figure 3D. They could use the "previous image" button to backup through the series. The small menu also contained a button that would allow readers to proceed to the next patient. The readers were fully aware that once they moved on to the next patient, they could not go back.

Readers

Twenty-four volunteer radiology residents and fellows from our Department of Radiology were recruited as readers; twelve in experiment 1 (minor added fractures) and twelve in experiment 2 (major added fractures). Data collection for the two groups was completely independent. The experience of the two groups of radiologists was matched in terms of years of experience and certification. The first experiment included four second-year residents, three third-year residents, two fourth-year residents, and three fellows. The second experiment included four second-year residents, four third-year residents, two fourth-year residents, and two fellows. Seventeen volunteer orthopedic surgery residents and fellows served as readers in experiment 3. Experience in years of the orthopedic surgery residents and fellows was matched with that of the radiology residents and fellows. The third experiment included four second-year residents, three third-year residents, four two fourth-year residents, three fifth-year residents and three fellows. All readers 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 reader read instructions and, with a demonstration case, was shown how to display the images, make responses, and advance through the cases. They were told that the purpose of the study was to better understand how radiologic studies are read so that error in interpretation can be reduced or eliminated. Readers were instructed to search for all acute fractures and dislocations and to identify each abnormality by placing the mouse cursor over the abnormality and clicking with the right mouse button. This produced a menu box for rating their confidence that the finding was truly abnormal. The readers were directed to indicate their confidence that a finding was abnormal by using discrete terms such as "definitely a fracture or dislocation", "probably a fracture or dislocation", "possibly a fracture or dislocation", and "probably not a fracture or dislocation, but some suspicion." These discrete terms were transformed into an ordinal scale where 1 represented no report, 2 represented suspicion, 3 represented possible abnormality, 4 represented probable abnormality and 5 represented definite abnormality. They were also directed to rate their confidence that the finding was abnormal by using a subjective probability scale from 0% to 100% using 20 categories (0%, 1–4%, 5%, 6–10%, 11–15%, 16–20%, 21–25%, 26–30%, 31–40%, 41–50%, 51–60%, 61–70%, 71–75%, 76–80%, 81–85%, 86–90%, 91–94%, 95%, 96–99%, 100%). The reason for the second scale was that more categories may sometimes yield more ROC operating

points, a potentially useful outcome for ROC curve fitting. A practice case followed to orientate each reader to the necessary mouse functions.

Data for each experiment were collected in two sessions separated in time by several months. Half of the patients presented in each session were from the SOS condition and the non-SOS (control) condition. Thus, in the course of the two sessions, each patient appeared twice, once in each experimental condition. Within each session, patients were presented in a pseudorandom order so that the occurrence of fractures was unexpected and balanced. Before each trial of an experiment, the reader was always informed of the patient's age and sex. During the reading sessions, viewing distance was flexible, room lights were dimmed to about 5 footcandles of ambient illumination, and there were no restrictions on viewing time.

The first examination in the series for each patient controlled the presence of the *added* fractures creating the non-SOS condition (without an added fracture) and the SOS condition (with an added fracture). For the first experiment using minor added fractures, the 70 added fractures were presented on digital radiographs of the foot [6], ankle [4], tibia/fibula [7], knee [5], pelvis [3], chest [3], shoulder [10], arm [5], elbow [4], wrist [11], and hand/fingers [12]. The digital radiographs that were used in place of these examinations for the non-SOS condition had the same distribution of body parts. For the second and third experiments using major added fractures, the 55 added fractures were presented on computed tomography of the cervical spine [48] or pelvis [7], and 15 of the added fractures were presented on digital radiography of the cervical spine [6], pelvis [7], chest [1] or hand [1].

Some of the radiographs appearing in the second or third examinations within each simulated patient presented a subtle *test* fracture; others presented no abnormalities. Whether the test fracture appeared in the second or third examination was random, but the same order was used for both experimental conditions in all experiments. The 47 test fractures were presented on digital radiographs of the foot [15], ankle [4], tibia/fibula [2], knee [3], shoulder [4], arm [3], elbow [1], wrist [3], and hand/fingers [8]. The 97 normal examinations appearing in the second or third positions of the series (47 with the test fractures, and 54 as pairs to make 27 normal patients) were presented on digital radiographs of the foot [7], ankle [13], tibia/fibula [3], knee [20], pelvis [19], chest [15], shoulder [3], arm [3], elbow [5], wrist [2], and hand/fingers [7].

ROC Analysis and Statistical Analysis

The multireader multicase (MRMC) ROC methodology developed by Dorfman, Berbaum and Metz (DBM) (11) has recently been extended in new software (DBM MRMC 2.1, available from http://perception.radiology.uiowa.edu) that allows the user to specify whether the analysis of variance model generalizes to readers, to patients, or both. Our primary method of analysis used the new DBM MRMC ROC analysis (11–16) fitting the discrete rating data with the contaminated binormal model, treating area under the ROC curve and sensitivity at specificity = 0.9 as measures of detection accuracy, and treating patients as a fixed factor, and readers as a random factor. The discrete scale was used because the subjective probability scale failed to provide additional operating points. The contaminated binormal model (17) was used because the rating data vector for test fractures was bimodal for every reader. One reader in the SOS condition of experiment 1 and another reader in the non-SOS condition of experiment 3 gave no false-positive responses. In this situation, ROC curves from the contaminated binormal model can differ in shape from those from data with false-positive responses. The use of sensitivity at particular value of specificity as the measure of diagnostic accuracy minimizes the effect of differing ROC curve shape. Specificity of 0.9 was chosen as a convenient level at which to measure sensitivity because it was the even value that maximized the number of readers with operating points on both sides of the sensitivity value. To be conservative, patients were treated as a fixed factor rather than random. While the original examinations included in this study were sampled from the population of patients with subtle fractures and/or

dislocations or from the population of patients without fractures, whether our simulated multitrauma patients—each of which was assembled from examinations of several patients— may not be a sample from an identifiable population. Because satisfaction of search affects readers rather than patients, generalization to the population of readers is fundamental. None of these analytic choices were critical to the results reported here; other choices—the subjective probability rating scale, other proper ROC models, treating cases as a random factor—lead to the same conclusions. As an additional check, we also report a rather traditional analysis in medical image perception research. Accuracy parameters were estimated by fitting the contaminated binomial model to the rating data of individual readers in each treatment condition and the ROC areas and sensitivities with and without the added lesions were compared using nonparametric Wilcoxon signed-rank tests (18,19).

An additional analysis was performed to compare the detectability of the fractures added within the three experiments. Discrete ratings of the first examination were treated as normal when no added fracture was present and as abnormal when an added fracture was present. So for each reader in each experiment there were 70 cases without added fractures and 70 with added fractures. Accuracy parameters were estimated by fitting the contaminated binomial model to the rating data of individual readers in each treatment condition and the ROC areas and sensitivities for detecting added fractures were compared between experiments 1 and 2, and between experiments 2 and 3 using relied on Mann-Whitney rank sum tests (18,19).

RESULTS

The following detailed results show that (1) an SOS effect demonstrated in the first experiment, with detection of subtle test fractures substantially reduced after low morbidity fractures were presented, (2) an SOS effect was not demonstrated in the second or third experiments, where added fractures were associated with high morbidity, and (3) low morbidity added fractures were slightly more detectable than high morbidity added fractures.

Experiment 1. Radiology Readers and Minor Added Fractures

As shown in Figure 4, for radiology readers, the DBM procedure demonstrated a reduction in ROC area for detecting subtle test fractures when a minor fracture was added to the first examination of the series (ROC area = 0.86 without added fracture vs. 0.81 with added minor fracture, difference = 0.05, $F(1,11) = 11.22$, $p < 0.01$). The DBM procedure also demonstrated a reduction in sensitivity at specificity = 0.9 for detecting subtle test fractures when a minor fracture was added to the first examination of the series (sensitivity = 0.71 without added fracture vs. 0.62 with minor added fracture, difference = 0.09, $F(1,11) = 8.33$, $p < 0.05$).

The supplementary analyses using Wilcoxon signed-rank tests performed on individual ROC parameters in each treatment confirmed the conclusions from the DBM MRMC analysis. A statistically significant reduction in detection accuracy for *test* fractures was found when the minor fractures were *added* (ROC area was 0.86 without minor added fracture vs. 0.81 with minor added fracture in the initial examination, $p < 0.01$; sensitivity at specificity of 0.9 was 0.71 without minor added fracture vs. 0.62 with minor added fracture in the initial examination, $p < 0.01$). The magnitude of the SOS effect is further characterized in the Appendix found at the end of this article.

Experiment 2. Radiology Residents and Fellows and Major Added Fractures

As shown in Figure 5, for radiology readers, the DBM procedure failed to demonstrate a difference in ROC area for detecting subtle test fractures with the inclusion of the initially presented distractor fracture (ROC area = 0.84 without major fracture vs. 0.84 with major fracture in the initial examination, difference = 0.0, $F(1,11) = 0.07$, $p = 0.94$). The DBM

procedure also failed to demonstrate a difference in sensitivity at specificity = 0.9 for detecting subtle test fractures with the inclusion of the initially presented distractor fracture (sensitivity = 0.66 without major fracture vs. 0.66 with major fracture in the initial examination, difference = 0.0, $F(1,11) = 0.01$, $p = 0.91$).

The supplementary analyses using a Wilcoxon signed-rank test performed on individual ROC parameters in each treatment confirmed the conclusions from the DBM MRMC analysis. For radiology readers with added major fractures, the test failed to demonstrate a statistically significant difference in detection accuracy for *test* fractures with *added* major fractures (ROC area = 0.84 without major fracture vs. 0.84 with major fracture in the initial examination, $p = 0.56$; sensitivity at specificity of 0.9 was 0.66 without minor fracture vs. 0.66 with minor fracture in the initial examination, $p = 0.94$).

Experiment 3. Orthopedic Surgery Residents and Fellows and Major Distractors

As shown in Figure 6, for orthopedic surgery readers, the DBM procedure failed to demonstrate a difference in ROC area for detecting subtle test fractures with the inclusion of the initially presented distractor fracture (ROC area = 0.81 without major fracture vs. 0.80 with major fracture in the initial examination, difference = 0.01, $F(1,16) = 0.31$, $p = 0.59$). The DBM procedure also failed to demonstrate a difference in sensitivity at specificity = 0.9 for detecting subtle test fractures with the inclusion of the initially presented distractor fracture (sensitivity = 0.61 without major fracture vs. 0.62 with major fracture in the initial examination, difference = −0.01, $F(1,16) = 0.08$, $p = 0.78$).

The supplementary analyses using a Wilcoxon signed-rank test performed on individual ROC parameters in each treatment confirmed the conclusions from the DBM MRMC analysis. For orthopedic surgery readers with added major fractures, the test failed to demonstrate a statistically significant difference in detection accuracy for *test* fractures with *added* major fractures (ROC area = 0.81 without major fracture vs. 0.81 with major fracture in the initial examination, $p = 0.59$; sensitivity at specificity of 0.9 was 0.61 without major fracture vs. 0.62 with major fracture in the initial examination, $p = 0.72$).

Detectability of Added Fractures

Results of the Mann-Whitney rank sum test comparing the first and second experiments demonstrated a statistically significant difference in detecting added fractures based on whether the fractures were associated with low or high morbidity (ROC area = 0.97 vs. 0.95, $p < 0.01$; sensitivity at specificity of 0.9 = 0.91 vs. 0.87, $p < 0.01$). A similar analysis performed on the raw number of added fractures reported at any level of certainty showed that minor fractures were reported more frequently than major fractures (69.6 vs. 66.8, $p < 0.001$).

Results of the Mann-Whitney rank sum test comparing the second and third experiments failed to demonstrate a statistically significant difference in detecting added fractures based on whether the readers were radiology residents or orthopedic surgery residents (ROC area = 0.95 vs. 0.95, p = 0.58; sensitivity at specificity of 0.9 =0.87 vs. 0.87, $p = 0.64$). A similar analysis performed on the raw number of added fractures reported at any level of certainty showed that radiology readers reported major fractures more frequently than did orthopedic surgery readers (66.8 vs. 63.1, $p < 0.001$). The lack of difference in accuracy parameters with a difference in frequency of true-positive reports suggests a difference in thresholds for radiology and orthopedic surgery readers (20).

DISCUSSION

Laboratory studies of satisfaction of search provide an operational definition of SOS that is lacking in retrospective and/or anecdotal accounts of errors. According to this definition, the abnormality missed because of the presence of another abnormality is also shown to be detected in the absence of the other abnormality. The "SOS effect" can only be verified when this definition holds, and where we measure detection accuracy in a way that is not altered when sensitivity is traded for compensatory changes in specificity. Only one laboratory study demonstrating the SOS effect in this way with radiographic studies of patients suffering from multiple injuries is available (2). Our first experiment replicated the SOS effect in multi-trauma patients with modern digital acquisition and display methods using new images and readers. Aside from the obvious value of replication, our first experiment provides an essential contrast to the absence of SOS in the experiments with high morbidity fractures. Were it not for this confirmation of an SOS effect using the same test images that failed to produce the SOS effect with high morbidity fractures, we might question whether the earlier finding of SOS in musculoskeletal radiology (2) was actually false positive. As it stands, we have two experiments with different readers and different simulated cases that demonstrate an SOS effect with low morbidity added fractures, and two experiments with different types of readers that fail to demonstrate an SOS effect with high morbidity added fractures.

Our second and third experiments tested whether the severity of an *added* fracture determines the magnitude of SOS reduction in detecting *test* fractures. This "severity hypothesis" came not from an ROC study, but a study measuring gaze time on missed fractures to show that the fractures were not missed because of faulty visual scanning (3). The technique of recording direction of gaze with high accuracy required the observer to wear a rather uncomfortable optical apparatus. Procedures to establish and check the correspondence between recorded and true gaze positions were required. These factors limited the number of examinations that could be read in a session and the willingness of readers to participate in multiple sessions. By contrast, ROC methodology requires more cases than ordinarily can be included in an eye-position study and many of these cases must be normal. The eye-position study included ten simulated multi-trauma patients, and nine of these included a test fracture (3). The presence of an SOS effect was only indicated by comparing the number of SOS events (where the test fracture was reported without an added fracture but missed when a fracture was added), with the number of anti-SOS events (where the test fracture was missed without an added fracture but reported when a fracture was added). With minor fractures, there was no difference in the number of SOS and anti-SOS false-negative responses; with major fractures, there were more SOS false-negative responses than anti-SOS false-negative responses. This evidence that detection of subsequent fractures is inversely related to severity of the initially detected fracture was considered tentative given the number of cases that were studied.

The experimental manipulation used to test the severity hypothesis was to substitute a CT or radiograph of the cervical spine or pelvis as the initial examinations of each case to present fractures associated with high morbidity. The inclusion of orthopedic surgery readers in the third experiment was predicated on the idea that injuries requiring immediate medical intervention may have a greater effect on orthopedic surgeons than on radiology consultants.

Surprisingly, high morbidity added fractures failed to demonstrate SOS effects. Why should this be so? To represent current practice, most major added fractures were shown with computed tomography. Is this the crucial difference? CT examinations provide much more image data to consider than a radiograph, but on the other hand, CT is used rather than radiography precisely because abnormalities are more clearly seen. Were the high morbidity added fractures more conspicuous than the low morbidity added fractures? If so, detection of the obvious added lesion might consume more perceptual resources so that more remain to

find the test lesion, thereby preventing SOS. Our analysis of the detectability of added fractures argues against this explanation: low morbidity added fractures were slightly more detectable by radiology readers than high morbidity added fractures.

Another possible explanation is that this was a laboratory study in which the observers knew that no clinical action was needed to follow up on findings of high morbidity lesions. Or perhaps in normal clinical practice, only senior radiologists are called on to decide on the need for clinical intervention. If residents and fellows were spared such judgments, then morbidity associated with a found abnormality might not affect them. Unfortunately, these explanations would predict no more SOS with high morbidity added fractures than with low morbidity added fractures, but they do not explain less SOS.

Unlike previous studies of SOS (2,3), the imaging modalities presenting the added and test fractures were different. This opens the possibility that CT eliminated SOS because SOS effects do not extend from one imaging modality to another. The anatomy presenting major vs. minor fractures differed. For the minor fractures, not only were the added and test fractures both presented on radiography, there was more overlap in the regions of the body examined. To present major fractures, new anatomic regions—the spine and pelvis—were introduced. Satisfaction of search could be mediated by the impression of having finished with an anatomic region based on finding an abnormality there. If so, there ought to be less satisfaction of search when abnormalities come from different anatomic regions because it should be easier to keep track of what has been evaluated and what has not. To study this "overlap hypothesis" we would need to be able to control the presence of fractures within the same examinations. New techniques to accomplish this are becoming available (21). (Of course, modern electronic communications allow imaging studies to be routed to available readers best qualified to read a particular study. In ordinary practice this should reduce the chance of SOS from one modality onto another.) In any case, before accepting the notion that there is no SOS effect with major added fractures on CT, we need to investigate the possibility that a severe injury on a CT can produce satisfaction of search for other injuries on that same CT examination. This, in turn, would shed light on whether a windfall in reduced SOS on subsequent radiography results from the shift from radiography to CT to evaluate spine injuries. In addition, we hope to perform a future experiment testing whether fractures with major morbidity on radiographs—specific fractures of the hip, shoulder, knee, etc.—yield an SOS effect in further radiographs in a series using ROC methodology.

We may suppose that the SOS effects found and not found in this study were mediated by the allocation of attention on different examinations of the patient's series. To discover whether reduced search explains an observed SOS effect in the first experiment and whether search is unchanged in the absence of an SOS effect in the second and third experiments, we conducted an analysis of viewing time on the various examinations and time to report fractures and false-positives. Addition of fracture, especially a serious fracture, increased the viewing time for the examination in which the added fracture appeared, but had little effect on viewing time for subsequent examinations in the patient's series or for time to make true- and false-positive reports. This is generally consistent with the results of the previous ROC study (2) in which time to report test fractures did not differ when they appeared with and without added fractures. By contrast, the previous eye-tracking study (3) showed a clear change in search behavior with less time devoted to subsequent examinations when a fracture was added to the first examination of the series, although the severity of the added fracture did not make much difference in viewing time. In the current experiments, the search of radiology readers was marked by longer viewing times for all examinations than orthopedic surgery readers and by somewhat slower report of false-positives. True-positive reports of test fractures were about as fast for orthopedic readers as radiology readers. Moreover, viewing time on radiographs with fractures missed only in the SOS condition did not differ from that of fractures missed

only in the non-SOS condition. None of these results suggest that global changes in visual search cause the SOS effect with minor added fractures.

References

1. Rogers LF, Hendrix RW. Evaluating the multiply injured patient radiologically. Orthopedic Clinics of North America 1990;21:437–447. [PubMed: 2367099]

2. Berbaum KS, El-Khoury GY, Franken EA Jr, Kuehn DM, Meis DM, Dorfman DD, Warnock NG, Thompson BH, Kao SCS, Kathol MH. Missed fractures resulting from satisfaction of search effect. Emergency Radiology 1994;1:242–249.

3. Berbaum KS, Brandser EA, Franken EA Jr, Dorfman DD, Caldwell RT, Krupinski EA. Gaze dwell times on acute trauma injuries missed because of satisfaction of search. Academic Radiology 2001;8:304–314. [PubMed: 11293778]

4. Rogers, LF. Radiology of Skeletal Trauma. New York: Churchill-Livingstone; 1982. p. 1

5. Rogers LF. Common oversights in the evaluation of the patient with multiple injuries. Skeletal Radiology 1984;12:103–111. [PubMed: 6484594]

6. LeBlang SD, Nunez DB Jr. Helical CT of cervical spine and soft tissue injuries of the neck. Radiologic Clinics of North America 1999;37:515–532. [PubMed: 10361544]

7. Nunez DB Jr, Quencer RM. The role of helical CT in the assessment of cervical spine injuries. American Journal of Roentgenology 1998;171:951–957. [PubMed: 9762974]

8. Berne JD, Velmahos GC, El-Tawil Q, Demetriades D, Asensio JA, Murray JA, Cornwell EE, Belzberg H, Berne TV. Value of complete cervical helical computed tomographic scanning in identifying cervical spine injury in the unevaluable blunt trauma patient with multiple injuries: A prospective study. Journal of Trauma-Injury Infection & Critical Care 1999;47:896–903.

9. Novelline RA, Rhea JT, Rao PM, Stuk JL. Helical CT in emergency radiology. Radiology 1999;213:321–339. [PubMed: 10551209]

10. Blackmore CC, Mann FA, Wilson AJ. Helical CT in the primary trauma evaluation of the cervical spine: an evidence-based approach. Skeletal Radiology 2000;29:632–639. [PubMed: 11201032]

11. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. Investigative Radiology 1992;27:723–731. [PubMed: 1399456]

12. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: Factorial experimental design. Academic Radiology 1998;5:591–602. [PubMed: 9750888]

13. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Academic Radiology 2004;11:1260–1273. [PubMed: 15561573]

14. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Statistics in Medicine 2005;24:1579–1607. [PubMed: 15685718]

15. Hillis SL. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Academic Radiology 2005;12:1534–1541. [PubMed: 16321742]

16. Hillis SL. A comparison of denominator degrees of freedom for multiple observer ROC analysis. Statistics in Medicine. 2006in press

17. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data. II. A formal model. Academic Radiology 2000;7:427–437. [PubMed: 10845402]

18. Dixon, WJ. BMDP Statistical Software Manual. 1. Berkekey, CA: University of Califormia Press; 1992. p. 155-174.p. 201-227.p. 521-564.

19. BMDP3D, BMDP7D, and BMDP2V, Release: 8.0. Copyright 1993 by BMDP Statistical Software, Inc. Statistical Solutions Ltd., 8 South Bank, Crosse's Green, Cork, Ireland, (http://www.statsol.ie).

20. Berbaum KS, Franken EA Jr, Dorfman DD, Miller EM, Krupinski EA, Kreinbring K, Caldwell RT, Lu CH. Cause of satisfaction of search effects in contrast studies of the abdomen. Academic Radiology 1996;3:815–826. [PubMed: 8923900]

21. Madsen MT, Berbaum KS, Ellingson AN, Thompson BH, Mullan BF, Caldwell RT. A new software tool for removing, storing and adding abnormalities to medical images for perception research studies. Academic Radiology 2006;13:305–312. [PubMed: 16488842]

22. McNemar, Q. Psychological Statistics. 4. New York: Wiley; 1969. p. 54-58.

Appendix: Quantifying the Magnitude of the SOS Effect

We quantify the magnitude of the satisfaction of search effect in the first experiment by treating the detection data as dichotomous categories: test fractures were either reported (at any level of certainty) or not reported. With use of these response categories, we classified the detection of test fractures into one of four categories, as follows: 1 = test fracture was missed with and without the added fracture; 2 = test fracture was reported without an added fracture but missed when a fracture was added ("SOS outcome"); 3 = test fracture was missed without an added fracture but reported when a fracture was added ("anti-SOS outcome"); and 4 = test fracture was reported both with and without the added fracture. Although not as compelling as ROC analysis, one can test for the presence of an SOS effect in the data by testing the null hypothesis that the population probabilities of outcome categories 2 and 3 are equal. We often do this in eyetracking experiments where much less detection data are available than in ROC experiments. The McNemar test for non-independent proportions (22) is inappropriate for counts summed over readers and cases because the counts are not identically distributed unless every reader has the same population probability of SOS. Nevertheless, we can evaluate SOS over readers. By using the four categories of native abnormality detection outcome defined earlier, we constructed a separate McNemar-type fourfold table for each of the readers in an experiment. We can test the McNemar null hypothesis for dependent sample frequencies taken over independent readers by using a Wilcoxon test, in which we generalize to the population of readers, but not to the population of cases. Under the McNemar null hypothesis, the population frequency of SOS events equals the population frequency of anti-SOS events. Because the study on minor distractors already demonstrated the SOS effect, a one-tailed test is the appropriate test on the frequencies.

Because a significant SOS effect was found, we further quantify the magnitude of the effect. The average number of SOS events per reader was 5.8, and the average number of anti-SOS events per reader was 4.2, for a sample mean difference of 1.6 ($p = 0.055$ one-tailed, marginally significant). This gives us a way to think about the size of the SOS effect for our sample: SOS is an increase 1.6 misses for 43 subtle fractures presented.
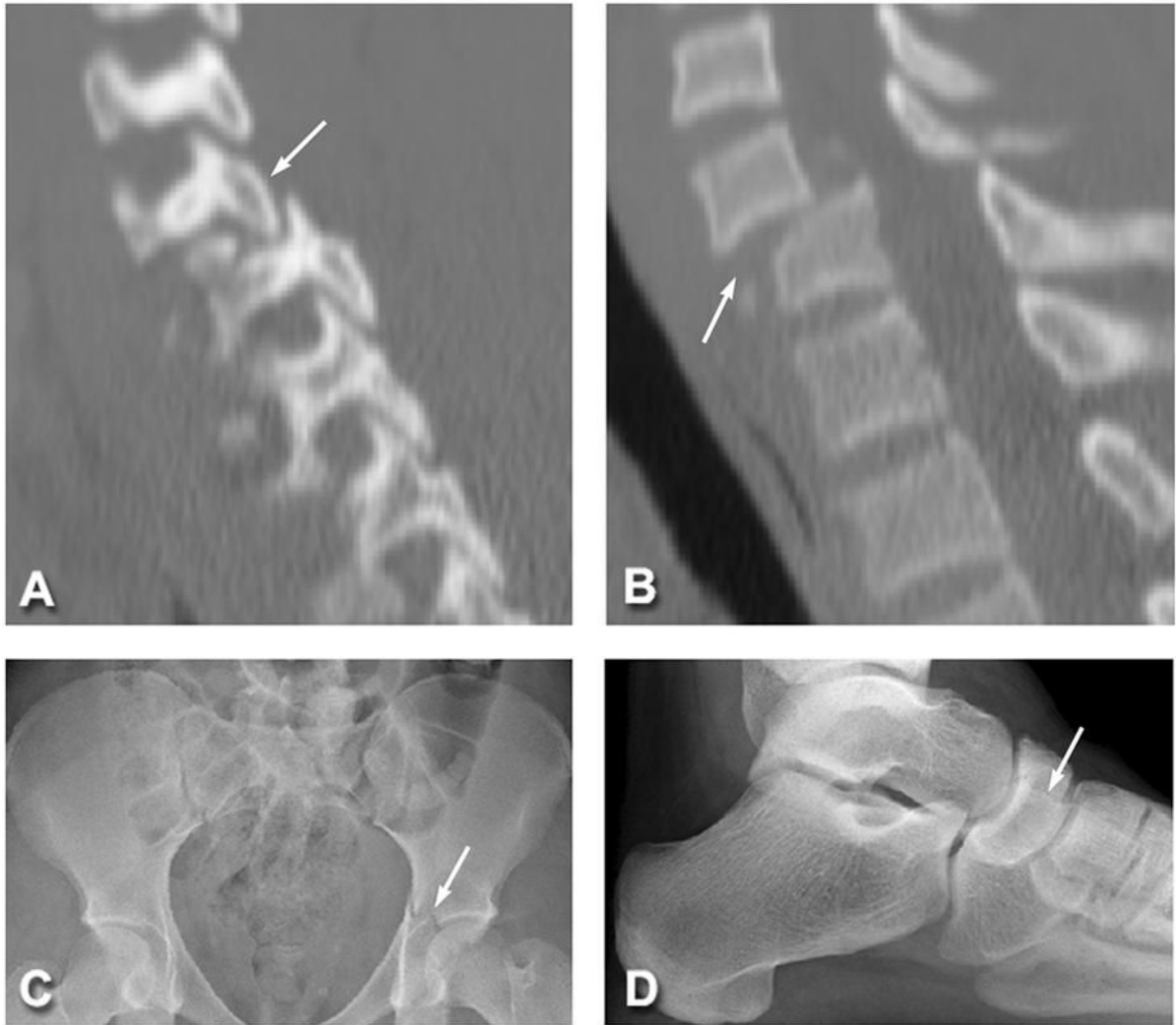
Figure 1.
The fractures used in case 28 in different experiments and experimental conditions. **A** and **B** represent obvious abnormalities with major clinical significance (arrows) and include a C5 bilateral facet dislocation **(A)** with anterior translation of C5 on C6 **(B)**. The acetabular fracture (arrow) in **C** represents a noticeable, but less significant abnormality. The navicular fracture (arrow), while the least obvious, is best seen on the lateral view **(D)**, although the AP and oblique views were also presented. The navicular fracture serves as a test fracture. Its detection was measured when a fracture of major clinical significance **(A** and **B)** was present in the case and when it was not. In a separate experiment, the detection of the target fracture **(D)** was measured when a more obvious, though non-life threatening fracture **(C)** was present in the case and when it was not.
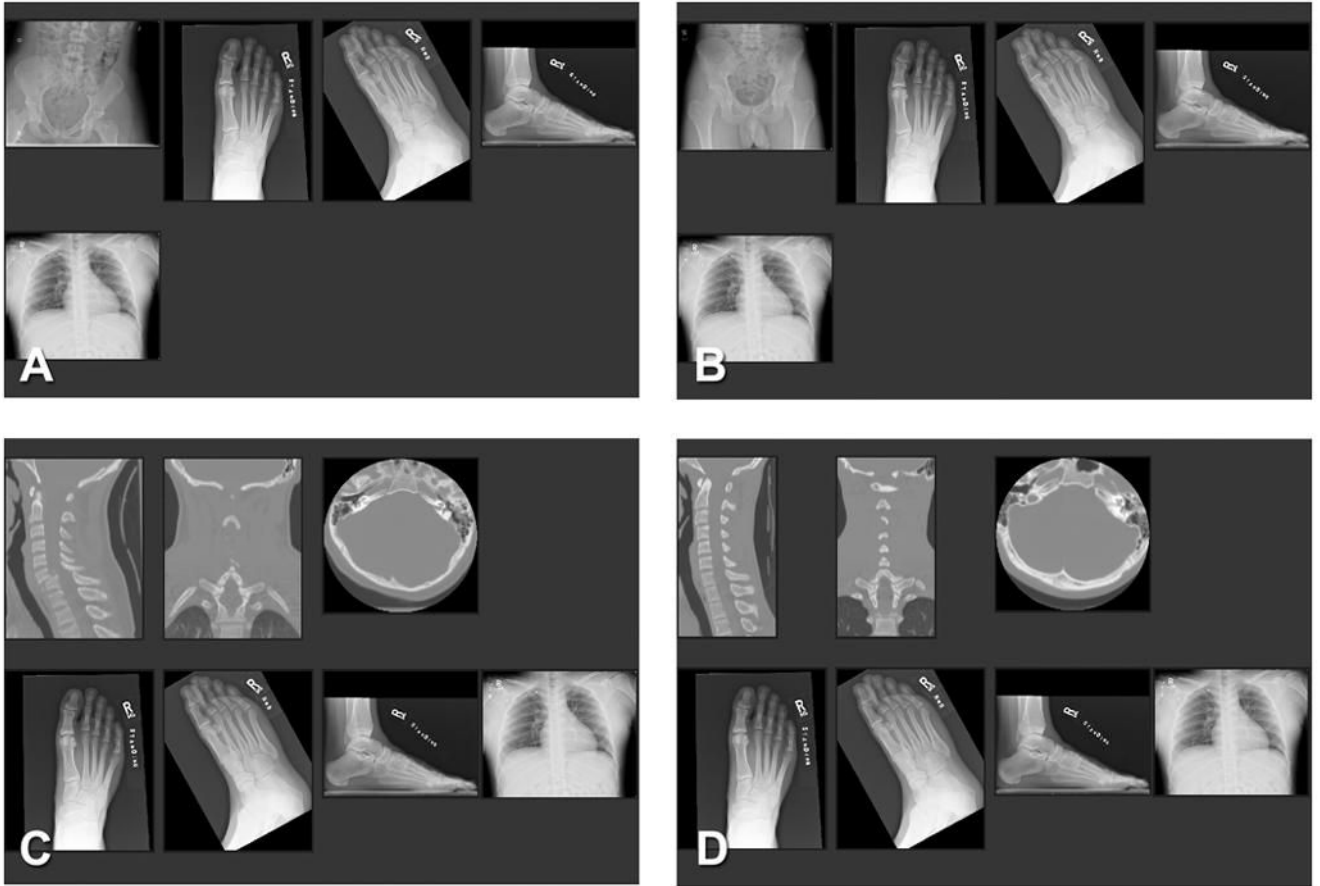
Figure 2.
Case 28 as it appeared in thumbnails in: **(A)** the SOS condition of experiment 1; **(B)** the non-SOS condition of experiment 1; **(C)** the SOS condition of experiment 2; and **(D)** the non-SOS condition of experiment 1. The test fracture was a right navicular fracture visible on the foot (when shown on a 3 megapixel monitor, **A–D**). The chest examination was normal **(A–D)**. The minor added fracture in experiment 1 was a left acetabular fracture **(A)**. The corresponding pelvic study for the non-SOS condition was a normal pelvis **(B)**. The major fracture in experiment 2 was a C6 anterior subluxation with bilateral facet fracture **(C)**. The corresponding cervical spine CT study for the non-SOS condition was normal **(D)**.
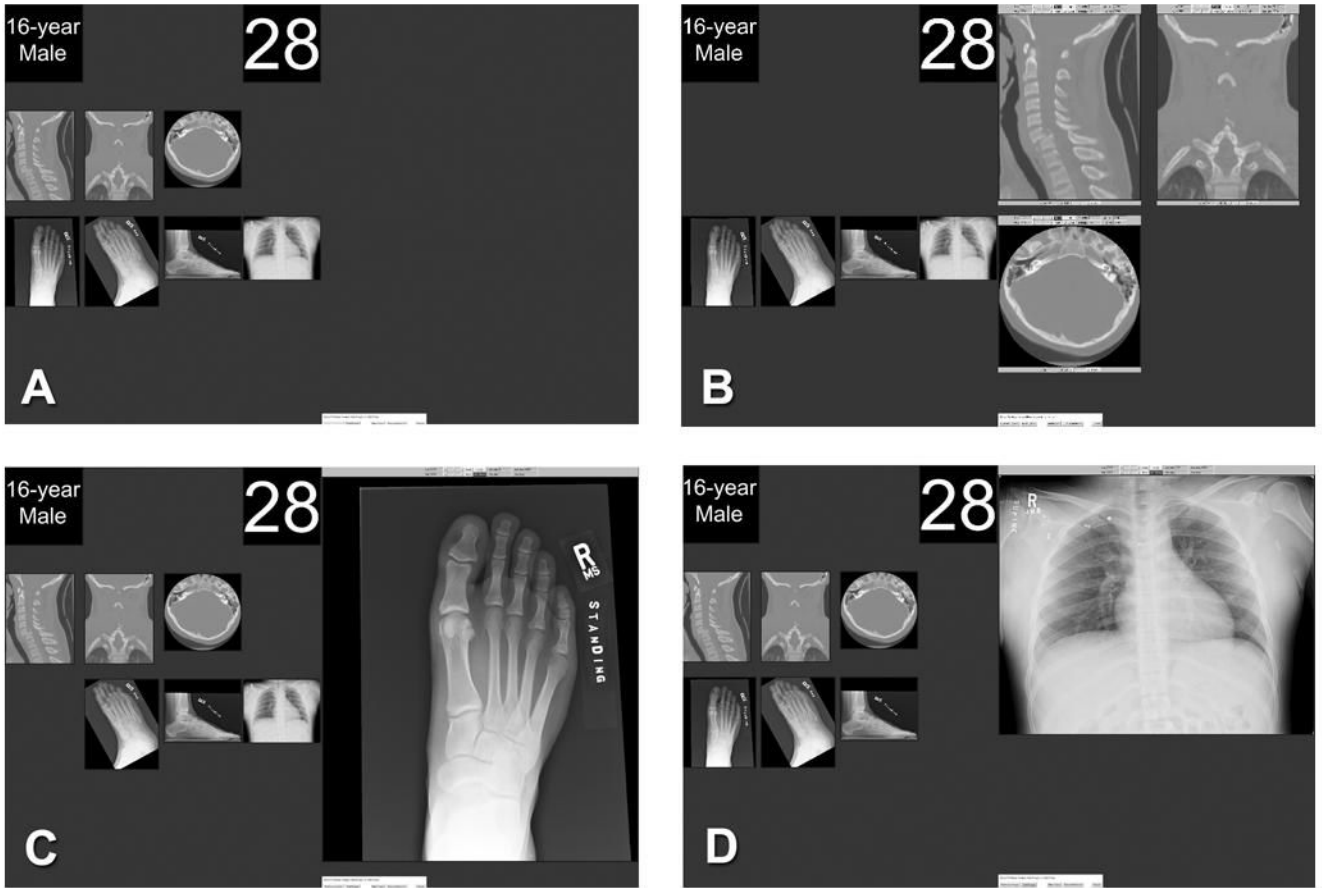
Figure 3.
A dual-monitor display of case 28 as it appeared in experiments 2 and 3. Each case was initially presented with patient information, case number, and thumbnail images comprising the complete study on the left monitor **(A)**. Images were displayed, one examination at a time, on the right monitor for diagnostic interpretation **(B–D)**. In the first examination display **(B)**, the three images are sagittal, coronal, and axial image stacks of a cervical spine CT demonstrating a C6 anterior subluxation with bilateral facet fracture. This fracture/subluxation is associated with major morbidity and served as a major *added* fracture. The next three images are digital radiographs of the foot demonstrating a subtle navicular fracture that served as a *test* fracture **(C)**. The chest radiograph did not demonstrate a fracture **(D)**.
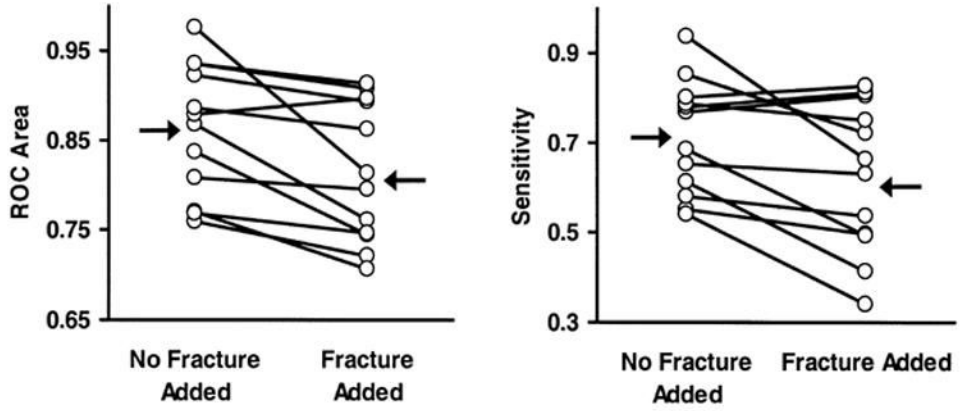
Figure 4.
Results of DBM MRMC analyses performed on the first experiment using contaminated binormal ROC area as the accuracy parameter (left panel) and sensitivity at specificity = 0.9 as the accuracy parameter (right panel). Each dot indicates an individual reader's accuracy; arrows indicate mean accuracy for the two conditions. An SOS effect was demonstrated.
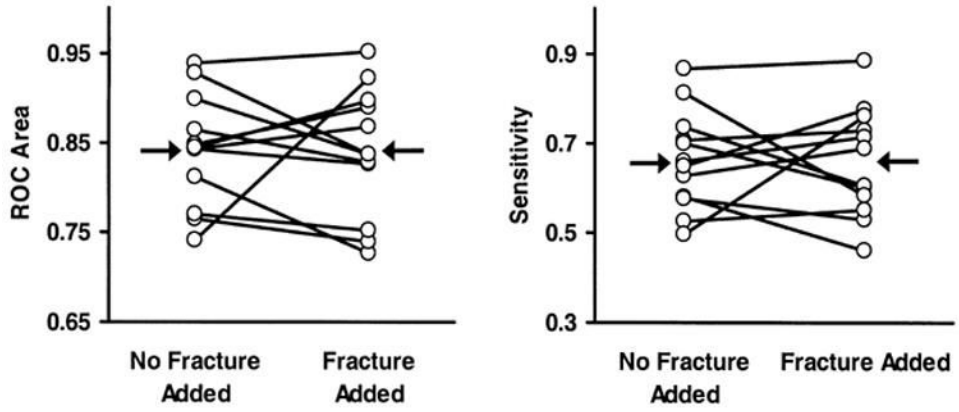
Figure 5.
Results of DBM MRMC analyses performed on the second experiment using contaminated binormal ROC area as the accuracy parameter (left panel) and sensitivity at specificity = 0.9 as the accuracy parameter (right panel). Each dot indicates an individual reader's accuracy; arrows indicate mean accuracy for the two conditions. No SOS effect was demonstrated.
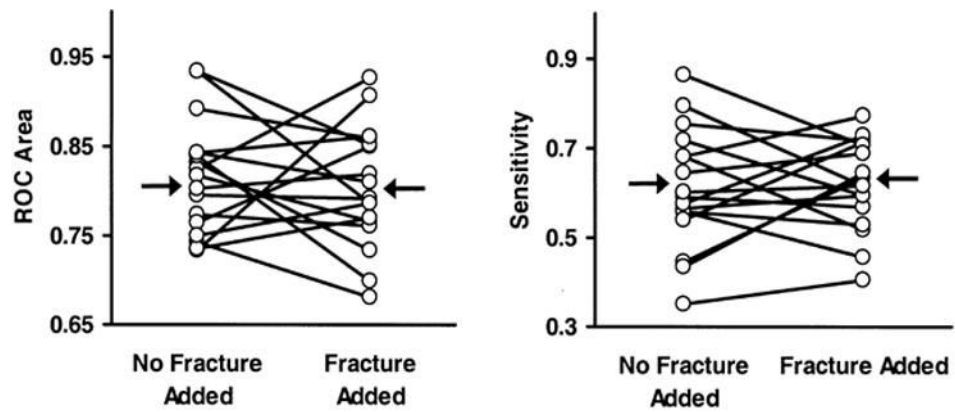
Figure 6.
Results of DBM MRMC analyses performed on the third experiment using contaminated binormal ROC area as the accuracy parameter (left panel) and sensitivity at specificity = 0.9 as the accuracy parameter (right panel). Each dot indicates an individual reader's accuracy; arrows indicate mean accuracy for the two conditions. No SOS effect was demonstrated.