Effect of fatigue on reading computed tomography examination of the multiply injured patient

Elizabeth A. Krupinski
Kevin M. Scharz
Mark S. Van Tassell
Mark T. Madsen
Robert T. Caldwell
Kevin S. Berbaum

Effect of fatigue on reading computed tomography examination of the multiply injured patient

Elizabeth A. Krupinski,^{a,*} Kevin M. Scharz,^b Mark S. Van Tassel,^b Mark T. Madsen,^b Robert T. Caldwell,^b and Kevin S. Berbaum^b

^aEmory University, Department of Radiology and Imaging Sciences, Atlanta, Georgia, United States

^bUniversity of Iowa, Roy J. and Lucille A. Carver College of Medicine, Department of Radiology, Iowa City, Iowa, United States

Abstract. Our goal was to ascertain how fatigue affects performance in reading computed tomography (CT) examinations of patients with multiple injuries. CT images with multiple fractures from a previous study of satisfaction of search (SOS) were read by radiologists after a day of clinical work. Performance in this study with fatigued readers was compared to a previous study in which readers were not fatigued. Detection accuracy for obvious injuries was not affected by fatigue, but accuracy for subtle fractures was reduced ($P = 0.016$). An SOS effect on decision thresholds was evident mirroring recent studies. Without fatigue, readers spent more time interpreting and reporting findings as the number of the injuries increased. When fatigued, readers did not increase reading time as fracture number increased. Without fractures, reading time for not-fatigued and fatigued readers was the same ($P = 0.493$) but was significant ($P = 0.016$) with an added subtle fracture. The difference increased with a major injury ($P = 0.003$) and increased further with both a major injury and subtle fracture ($P = 0.0007$). Fatigue and multiple abnormalities have independent effects on detection performance but do interact in determining search time. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.4.3.035504](https://doi.org/10.1117/1.JMI.4.3.035504)]

Keywords: fatigue; errors; interpretation time; satisfaction of search.

Paper 17183RR received Jun. 20, 2017; accepted for publication Sep. 12, 2017; published online Sep. 29, 2017.

1 Introduction

There has been growing concern about the impact of fatigue on performance in radiology, specifically with respect to increased error rates, physical injury, mental and emotional stress, and burnout.^{1–4} Increased imaging has increased workload and led to longer hours interpreting cases.² The effects of an extended workday have been investigated in laboratory studies using a number of tasks that simulate clinical interpretation.^{5–7} Studies using single lesion detection tasks [fractures in bone images and nodules in computed radiography (CR) and computed tomography (CT) chest] generally demonstrated a statistically significant average decrease in diagnostic accuracy of ~4%.^{5,6} The first study investigating the effects of fatigue on interpretation involving multiple abnormalities was recently published.⁷ This study, however, was limited to plain film chest radiography thus the results may not generalize to more advanced imaging techniques, such as CT. In fact, earlier studies [that did not include satisfaction of search (SOS)] demonstrated that the impact of fatigue may indeed differ for plain film and chest CT so there is a reason to believe that the same may hold true with SOS.

For radiologists, visual fatigue and eye strain may be of particular importance. Physical measures of visual strain (accommodation and convergence) indicate that fatigue can reduce a radiologist's ability to focus, especially at close viewing distances.^{5,6} Subjective feelings of physical, mental, and emotional strain^{5,6} increase in radiologists who have spent as little as 8 h a day interpreting clinical images. Both physical and subjective measures indicate that residents may actually be more vulnerable to effects of fatigue than attending radiologists.

The earlier studies^{5,6} tested only simple detection accuracy using solitary targets. The presence of multiple abnormalities in radiology examinations yields more reading errors than single abnormalities (see Ref. 8 for a historical review of the SOS phenomenon). Recent studies^{9,10} examining the SOS effect in chest radiography with modern imaging techniques indicate a threshold shift [reduction in both true positive (TP) and false positive (FP)], reflecting a reluctance to report native lesions in the presence of nodules. Using the same computed radiography examinations, Krupinski et al.⁷ were the first to study performance in detecting multiple abnormalities under conditions of fatigue.

The SOS effect they observed was similar to that of the recent SOS studies.^{9,10} SOS manifested as a reduction in willingness to respond (threshold shift). The center of the FP range was significantly reduced (from 0.10 to 0.05, $P < 0.001$), as was the center of the TP range (from 0.39 to 0.33, $P < 0.01$). Although the presence of nodules did not significantly affect inspection time, presence of a native abnormality did increase inspection time (from 40 to 55 s, $P < 0.0001$). The results suggest that fatigue does not fundamentally change the nature of the SOS effect. Krupinski et al.⁷ proposed a “working hypothesis” that the effect of fatigue (reducing TPs more than FPs) may simply add to those of the SOS effect (reducing both TPs and FPs). One goal of the current study is to test this working hypothesis more directly than was possible in the earlier study.⁷

Another goal is to further the study of cognitive fatigue in radiology. Fatigue may degrade such basic functions as short-term memory and other working memory systems, as well as interfere with attentional control and executive functions involved in decision-making. Such cognitive consequences can occur with or without feelings of sleepiness. In the latter

*Address all correspondence to: Elizabeth A. Krupinski, E-mail: ekrupin@emory.edu

case, readers may not fully appreciate the extent of their impairment.

To accomplish these goals, we replicated a recent study¹¹ that tested whether severe distracting fractures controlled the magnitude of SOS on other fractures when both appear in a single CT image. In other words, this study used the same set of test cases as in a multiple trauma SOS study using similar readers. However, we performed our experiment toward the end of a day of clinical reading when readers were fatigued rather than early in the day as in the previous study.

2 Methods and Materials

The study used the same general methods as the multiple trauma CT study,¹¹ which are explained here briefly. Studies assessing SOS in laboratory experiments use abnormalities defined as “test” abnormalities because their detection is measured. These test abnormalities are always presented twice: once with and once without another abnormality (added by the experimenter) in that same exam. When the test abnormality is missed in the presence (but not absence) of a distracting abnormality, SOS occurs.

2.1 Images

This study was approved by the University of Iowa Institutional Review Board (IRB). All images were stripped of all patient identifiers and given a unique code. Two senior skeletal radiologists with 15 and 37 years of experience verified the truth status of the cases. The cases were CT cervical spine exams (see Fig. 1 for an example), half with serious injuries to the cervical spine (SOS condition) and half with no apparent spinal injuries (non-SOS condition): 35 had high morbidity fractures (Table 1) and 35 had no fractures. The cases were paired to create 35 dyads matched in age and general physical characteristics. Sixteen dyads had subtle test fractures inserted into both exams of the pair (Table 1). For each exam, the same test fracture was simulated (for details, see Ref. 11). The other 19 dyads were used to gather FP reports for a receiver operating characteristic (ROC) curve analysis.

Table 1 Types (# cases) of high morbidity and simulated test fractures in study.

High morbidity fracture	Simulated test fractures
Burst (1)	Transverse (1)
Dens (2)	Lamina (1)
Facet (6)	Rib (4)
Facet dislocation (2)	Transverse process (1)
Pedicle (5)	Clavicle (2)
Lamina (2)	Spinous process (3)
Lateral mass (2)	Anterior arch (2)
Facet-lamina rib (3)	Scapula (1)
Spinous process (3)	Transverse foramen (1)
Bilateral arch (1)	—
Vertebral body-lamina (2)	—
Jefferson (2)	—
Superior facet (2)	—
Burst-lamina (1)	—

The experimental (SOS) condition included 35 cervical spine CT cases, all of which contained severe cervical spine injuries. For each of these 35 cases, a similar case was found that had no injuries. Image modification software was used to add simulated fractures to each pair of cases, with and without a major injury. Sixteen different small fractures were added to 16 of the 35 pairs of images. The 35 cases without native injuries constituted a control (non-SOS) condition. This process yielded a case set containing: 70 total cases, 16 with serious fracture and simulated fracture, 19 with just serious fracture 16 with just simulated fracture, and 19 with no fracture.

Fig. 1 Typical case pair showing a slice from (a) a patient with a serious fracture indicated by the white arrow and (b) a matched patient without a fracture in a similar location.

2.2 Fracture Simulation

This experiment and the previous one¹¹ required simulation of subtle test fractures on spine CT. Placing the test fracture in the same location of two different patients required expert guidance. Each pair was matched by patient age, general physical characteristics, and number of slices in the study, and both studies needed to share the same size, shape, intensity, and contrast.

A medical physicist and coauthor developed software to place simulated fractures in nearly identical locations of two different images. The software allowed the operator to draw a fracture region on the bone image using the computer mouse (see online Appendix of Ref. 11 for a detailed description). Bony structures limited the location of simulated test fractures that were deemed realistic by our musculoskeletal radiologists.

Figure 1 shows a typical case pair, showing a slice from a patient with a serious fracture [Fig. 1(a), indicated by the white arrows]. A matched patient without a fracture [Fig. 1(b)] appears in a similar location. Figure 2 shows other locations in these examinations, in which a simulated rib fracture (the target) appears in nearly identical locations of the paired patients as shown by white arrows.

2.3 Image Presentation

The images were displayed using a Dell Precision 360 Mini-tower and two 3-megapixel LCD monitors (National Display Systems, San Jose, California) calibrated to the Digital Imaging and Communications in Medicine Grayscale Standard Display Function. “WorkstationJ” display software was used to perform the reading study and collect observer responses, including location of abnormality and confidence rating for the abnormality.¹² Only the axial views of both the test and distracting abnormalities were presented. Room lights were set to 48 lux of ambient illumination, which is within the range of the American College of Radiology guidelines of 20 to 40 lux or brighter as may be tolerable or even desirable so long as conformance with the American Association of Physicists in Medicine Task Group 18 specifications is maintained.¹³

2.4 Readers

Twenty radiologists served as readers and signed an IRB-approved informed consent document. Readers were matched to those of the previous experiments in terms of experience and certification, with 6 second-year residents, 8 third-year, 4 fourth-year, and 2 fellows (total $n = 20$ readers). They participated after engaging in at least 8 h that day/night of clinical image interpretation.

In our previous studies,^{5–7} we measured visual fatigue with a refractometer (which measures accommodation or the ability to focus) and physical/mental fatigue with the Swedish Occupational Fatigue Inventory (SOFI). In all prior studies, all of the observers demonstrated statistically significant reductions in their ability to focus and increases in the SOFI measures of physical exertion, physical discomfort, sleepiness, lack of energy, and lack of motivation. Given the time involved in this study, we did not feel it necessary to acquire these data again.

2.5 Procedure

Prior to the study, readers were provided with a set of written instructions and were shown how to use the display program to view images and record their responses. They were told that their task was 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 triggered a response box to appear where the readers indicated their confidence (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%) that a fracture or dislocation was present.

Readers participated in two sessions separated by 2 to 3 months. In each session, half of the cases were from the SOS and half from the non-SOS condition, so over both sessions each case appeared twice, once in the experimental and once in the control condition. Cases appeared in a pseudorandom order so the occurrence of fractures was unexpected and balanced within each session. The readers were informed of the patient’s age and sex before each case presentation.

Fig. 2 Other locations in exams in which a simulated rib fracture (the target) appears in nearly identical locations of the (a) and (b) paired patients as shown by white arrows.

2.6 Statistical Analyses

2.6.1 Analysis of variance and statistical interaction

Krupinski et al.⁷ proposed a working hypothesis, in which the effect of fatigue may simply add to those of the SOS effect. They could not test this working hypothesis directly because their experiment with tired readers could not be directly compared to a previous one with rested readers.⁹ To test this working hypothesis directly, we replicated all aspects of a recent study with rested readers.¹¹

To combine the data of two experiments, each with a “within-subject” (reader) factor for the SOS manipulation, we used analysis of variance (ANOVA) in which the factor that distinguishes the two experiments, rested versus tired readers, is treated as a “between-subject” factor because the rested and tired readers were not the same individuals. The analysis and interpretation of interaction effects of orthogonally varied factors in the ANOVA are complex topics.¹⁴ Interaction effects represent the combined effects of factors, such as fatigue and the addition of an abnormality, on the dependent measure, such as detection accuracy, decision threshold, or inspection time. When an interaction effect occurs, the impact of one factor depends on the level of the other factor. In other words, an interaction occurs when the effect of factor *A* varies across levels of factor *B* (i.e., the combined effect of factors *A* and *B* is not simply additive). An advantage of ANOVA lies in its ability to estimate and test interaction effects. A milestone in statistical thinking was the recognition that multiple effects should be studied rather than simply the isolated effects of single variables.¹⁵ When interaction effects are present, the interpretation of the main effects of individual variables may be incomplete and misleading.

The ANOVA and statistical interaction are needed to study how fatigue and SOS interact to place added demands on the reader. To simplify elements of that analysis for readers who are not familiar with ANOVA, we will present simpler analytic results where possible.

2.6.2 Detection accuracy

Detection accuracy for test fractures in the non-SOS and SOS conditions was measured and compared to determine whether high morbidity fractures on spine CTs result in other fractures being missed. Detection accuracy was measured by scoring responses on simulated test fractures.¹¹ The multireader multicase (MRMC) ROC methodology developed by Dorfman, Berbaum, and Metz (DBM),^{16,17} and recently extended in software (OR-DBM MRMC 2.5¹⁸), was used to analyze the confidence ratings. As in the previous experiment, we used ROC analysis fitting the rating data with the contaminated binormal model (CBM),¹⁹ treating the area under the ROC curve (AUC) as the measure of detection accuracy. We also studied average sensitivity measured at the specificity of 0.9 because this index focuses on a part of the ROC curve well-supported with empirical ROC points. SOS affects readers rather than patients so we generalize to the population of readers, with patients treated as a fixed factor and readers as a random factor. These analytic choices mirrored those of previous studies.

Although OR-DBM MRMC 2.5 allowed us to compare the non-SOS and SOS conditions that software cannot currently be used to study the difference between nonfatigued and fatigued readers at the same time. Therefore, we continued our analysis

using the AUC and sensitivity measures produced by OR-DBM MRMC but subjected them to more complex ANOVA.

Although these experiments were designed so that the major injuries were easy to see, we also checked whether fatigued readers missed them more often. To compare the detection of not-fatigued and fatigued reading, we used the following scoring of major fractures. The 35 examinations of the SOS condition each included a major injury, and the confidence ratings pertaining to these fractures were treated as abnormal trials for ROC analysis. The 35 cases of the non-SOS condition had no major injuries, and the most abnormal falsely positive confidence ratings on these examinations were treated as normal trials for ROC analysis. Responses pertaining to test fractures were ignored for this analysis. The data for each reader in each experiment were fitted using the CBM and components of OR-DBM MRMC 2.5. Detection accuracy was then analyzed using two-group *t*-tests.²⁰

2.6.3 Decision thresholds

We studied the possibility that decision thresholds might shift in the presence of the major added injury. Even with no difference in AUC, sensitivity and specificity can vary inversely indicating shifts in decision thresholds. We studied decision thresholds using the FP fraction (FPF) associated with each ROC point, perhaps the best and simplest index of decision threshold. The FPFs were established by grouping the confidence ratings as 10%/20% to 100%, 10% to 20%/30% to 100%, 10% to 30%/40% to 100%, 10% to 40%/50% to 100%, 10% to 50%/60% to 100%, 10% to 60%/70% to 100%, 10% to 70%/80% to 100%, 10% to 80%/90% to 100%, and 10% to 90%/100%. For each reader-treatment combination, the midpoint of the FP range of ROC points was computed as the average of the most and least conservative ROC points. As such, for each experiment, we get an array of midpoint values consisting of the two SOS conditions $\times$ 20 readers (40 values). Likewise, the width of the FP range of ROC points was computed as least conservative minus the most conservative ROC point. We also studied the least conservative ROC point. We begin with a preliminary analysis testing for differences in FPFs between non-SOS and SOS conditions using *t*-tests. We then complete the analysis using ANOVA with a between-subject factor for rested versus tired and a within-subject factor for SOS manipulation.

2.6.4 Response times

We computed median response times for examinations that (1) had no abnormalities, (2) had only test fractures, (3) had only major injuries, and (4) had both test fractures and major abnormalities. In a preliminary analysis, we used two-group *t*-tests to compare nonfatigued and fatigued readers for each of these examination types. Next, we present a more complex analysis of response times using an ANOVA, which treated the four conditions listed already as crossed within-subject factors for test fracture presence and major injury presents and adds a between-subject factor for level of training.²¹

3 Results

3.1 Detection of Test Fractures

As guidance for the reader, we provide a short summary, here, in the hope that it will be easier to follow the rest. Detection accuracy for serious injuries was not affected by fatigue, but

Fig. 3 Average ROC points (left full ROC space, right insert close up of region of interest) for non-SOS (open symbols) and SOS conditions (solid symbols) of current (fatigued = triangles) and previous¹¹ (not fatigued = circles) experiments. Each point is an average of 20 readers.

detection accuracy for subtle fractures was reduced. For both experiments with rested and tired readers together, there was no statistically significant effect of the SOS manipulation on detection accuracy. There was no statistically significant interaction of SOS treatment by fatigue with regard to detection accuracy.

Figure 3 presents average ROC points across readers with empirical line segments connecting those points. The circles are from the previous (nonfatigued) experiment¹¹ and the triangles are from the current (fatigued) experiment. Although there is an apparent reduction in performance in the SOS condition relative to the non-SOS condition of that experiment, the OR-DBM procedure demonstrated only a modest and non-statistically significant reduction in ROC area for detecting subtle test fractures when a major injury was present in the CT examination [ROC AUC = 0.880 without major injury versus 0.855 with major injury; difference = 0.025; $F(1,19) = 1.95$, $P(\text{one-tailed}) = 0.089$]. The results for sensitivity at specificity of 0.9 were slightly more significant ($P = 0.07$). In any case, the SOS effect on detection accuracy was meager at best.

Figure 3 also suggests no SOS effect on detection accuracy for the current experiment as shown by the average ROC points indicated with triangles. This was born out by the results of the OR-DBM procedure applied to the data of the current experiment with tired readers. There was no reduction in ROC area for detecting subtle test fractures when a major injury was present in the CT examination [ROC AUC = 0.835 without major injury versus 0.833 with major injury; difference = 0.002; $F(1,19) = 0.02$, $P(\text{one-tailed}) = 0.448$]. The results for sensitivity at specificity of 0.9 were also not significant ($P = 0.227$).

We extended our analysis of detection of test fractures using an ANOVA that treated ROC AUC as the dependent variable and included a factor for experiment (the previous not-fatigued¹¹ and current fatigued studies) as a between-subject-independent variable. The SOS manipulation (absence or presence of major injury) was a within-subject-independent variable. Level of training (second-, third-, fourth-year resident, or fellow) was a between-subject-independent variable. There was a significant main effect of experiment [$F(1,31) = 6.56$, $P = 0.0155$] with fatigued performance (mean = 0.828) being lower than not fatigued (mean = 0.878) overall. Although performance increased as a function of year (second mean = 0.81; third = 0.85; fourth = 0.88; or fellow = 0.88), this effect was not

statistically significant [$F(3,31) = 2.36$, $P = 0.091$]. Although performance was lower in the SOS condition (mean = 0.846) than the non-SOS condition (mean = 0.859), the difference was not significant [$F(1,31) = 0.83$, $P = 0.369$]. The interaction of SOS manipulation by fatigue was not statistically significant [$F(1,31) = 0.73$, $P = 0.3981$].

A similar second analysis used sensitivity at a specificity of 0.9 as the dependent variable. The results were essentially the same, with experiment being significant [$F(1,31) = 2.76$, $P = 0.0217$, nonfatigued average sensitivity of 0.743 versus fatigued average sensitivity of 0.651] and year [$F(3,31) = 2.36$, $P = 0.091$] and SOS [$F(1,31) = 0.44$, $P = 0.512$] being not significant. The interaction of SOS manipulation by fatigue was not statistically significant [$F(1,31) = 0.02$, $P = 0.9011$].

3.2 Detection of Major (Distractor) Fractures

The average ROC AUC was 0.944 for readers who were not fatigued and 0.945 for readers who were [t(38) = -0.01, $P = 0.99$]. The average sensitivity at specificity of 0.9 was 0.896 for readers who were not fatigued and 0.892 for readers who were [t(38) = 0.14, $P = 0.89$]. Fatigue had no effect on the detection of these major injuries.

3.3 Decision Thresholds

The most important finding in this section is that, for subtle fractures, the main effect of SOS treatment was statistically significant on decision thresholds so that the range of the FPF was narrowed by the SOS treatment in both experiments. In the previous paper,¹¹ there was no statistically significant difference in either the average midpoint of FP points (0.13 versus 0.13, $P = 0.98$) or for the range of FP points (0.12 versus 0.10, $P = 0.32$). Figure 3 shows a difference in the range of FPFs associated with ROC points for the SOS in the current experiment where the readers were fatigued. There was no statistically significant difference in the midpoint of the range of FPFs associated with ROC points between non-SOS and SOS conditions [FPF = 0.099 for non-SOS versus 0.082 for SOS condition, difference = 0.017, $T(19) = 0.98$, and $P(\text{two-tailed}) = 0.337$]. However, the width of the range of FPFs associated with ROC points was significantly narrower for the SOS condition than the non-SOS condition [FPF = 0.124 for non-SOS versus 0.063 for SOS condition, difference = 0.061, $T(19) = 2.41$, $P(\text{two-tailed}) = 0.026$]. Finally, the FPF associated with the least conservative ROC point was only marginally less conservative in the SOS condition relative to the non-SOS condition [0.1605 versus 0.1132, difference = 0.0474, $T(19) = 1.74$, $P(\text{two-tailed}) = 0.0982$]. This suggests that the narrowing of the range of FPs associated with the ROC points in the SOS condition was based on less conservatism for the most lenient ROC points and greater conservatism for the least lenient ROC points.

The aforementioned t -tests are presented in part for simplicity and in part for continuity with the analysis presented in the previous paper.¹¹ ANOVA allows for both the previous experiment with rested readers and the current one with tired readers to be studied within a single analysis. For the midpoint of the range of FPFs associated with ROC points, the main effect of experiment (rested versus tired) was not statistically significant [FPF = 0.081 versus 0.090, $F(1,38) = 0.23$, $P = 0.6369$] nor was the main effect of SOS treatment [non-SOS versus SOS, 0.090 versus 0.082, $F(1,38) = 0.46$, $P = 0.5030$]. The

interaction of experiment by SOS was also not significant [$F(1,38) = 0.62$, $P = 0.4357$]. For the width of range of FPFs, the main effect of experiment (rested versus tired) was not statistically significant [0.107 versus 0.093, $F(1,38) = 0.19$, and $P = 0.6692$], but the main effect of SOS treatment was statistically significant [non-SOS versus SOS, 0.118 versus 0.082, $F(1,38) = 4.59$, and $P = 0.0386$]. The interaction of experiment by SOS was not significant [$F(1,38) = 1.90$, $P = 0.1764$]. This analysis suggests that although analysis of the previous experiment alone suggested no decision threshold effects, the current analysis shows that the range of the FPF was also narrowed by the SOS treatment in that experiment.

An analysis of the decision thresholds for reporting the major fractures showed that they were not affected by the readers' state of tiredness [center of the FP range: 0.045 versus 0.054, $F(1,38) = 0.45$, $P = 0.5048$; width of the FP range: 0.061 versus 0.069, $F(1,38) = 0.10$, and $P = 0.7543$].

3.4 Response Times

3.4.1 Preliminary analysis

Figure 4 shows average median response times for examinations that (1) had no abnormalities, (2) had only test fractures, (3) had only major injuries, and (4) had both test fractures and major abnormalities. Table 2 presents two-group t -tests that compare nonfatigued and fatigued readers for examination with increasing numbers of abnormalities. Both Fig. 4 and Table 2

demonstrate that when readers are not fatigued, they spend more and more time interpreting and reporting findings as the number and seriousness of the injuries increases. However, when readers are fatigued, there is a flattening of reading time even when there are more abnormalities to report.

3.4.2 Analysis of variance of response time

Median inspection times were computed for each reader for non-SOS and SOS conditions for examinations with and without test abnormalities. These four median times per reader were analyzed with an ANOVA with within-subject factors for SOS condition and examination type (with or without test abnormality). The main effect of fatigue was statistically significant [110 s for not tired versus 82 s for tired for a reduction of 28 s average inspection time, $F(1,38) = 7.90$, $P = 0.0078$]. The main effect of the SOS manipulation was statistically significant [82 s without an added serious fracture versus 109 s with a serious fracture added for an increase of 27 s, $F(1,38) = 85.27$, $P < 0.0001$]. The main effect of test fracture presence was statistically significant [91 s without test fracture versus 101 s with test fracture for an increase of 10 s, $F(1,38) = 32.55$, $P < 0.0001$]. Each of the two-way interactions of these factors was statistically significant, but the three-way interaction of experiment, SOS treatment, and case type was not.

There was a statistically significant interaction of the presence of a test fracture by the presence of a major fracture [$F(1,38) = 8.88$, $P = 0.0050$] (Table 3). The increase in

Fig. 4 Average across examinations of inspection time (seconds) in conditions with and without test fractures and with and without major injuries of current (fatigued = bars with lines) and previous (not fatigued = solid bars) experiments.¹¹ Each point is average across readers of median of cases within the four abnormality combinations.

Table 2 Seconds of reading time.

Test fracture	Major fracture	Readers not tired	Tired readers	Difference	P value [$T(38)$]
Absent	Absent	82.8	76.6	-6.2	0.4925
Present	Absent	95.8	74.7	-21.1	0.0155
Absent	Present	119.8	85.4	-34.4	0.0027
Present	Present	141.5	91.2	-50.3	0.0007

Table 3 Interaction effect of SOS treatment with presence of test fracture on reading time.

SOS condition (presence of major fracture)	Examinations without a test fracture (s)	Examinations containing a test fracture (s)	Effect of presence of a test fracture on reading time (s)
Non-SOS condition (no added major fracture)	79.7	85.2	+5.5
SOS condition (added major fracture)	102.6	116.4	+13.8
Effect of presence of major fracture on reading time	+22.9	+31.2	—

Table 4 Interaction effect of fatigue with SOS treatment on reading time.

Experiment	Non-SOS condition (no added major fracture) (s)	SOS condition (added major fracture) (s)	Effect of SOS manipulation on reading time (s)
Not tired ¹¹	89.3	130.6	+41.3
Tired (current experiment)	75.6	88.3	+12.7
Effect of fatigue on reading time	-13.7	-42.3	—

reading time when the test fracture was present rather than absent was far less when no other fracture was present than when the major fracture was also present (+6 s versus +14 s). Stated differently, the increase in reading time when a major fracture was added was far greater when the test fracture was present than absent (+31 s versus +23 s).

There was a statistically significant interaction of fatigue (experiment) by SOS condition, [$F(1,38) = 24.01$, $P = 0.0001$]

Table 5 Interaction effect of fatigue with presence of test fracture on reading time.

Experiment	Examinations without a test fracture (s)	Examinations containing a test fracture (s)	Effect of presence of a test fracture on reading time (s)
Not tired ¹¹	101.3	118.6	+17.3
Tired (current experiment)	81.0	83.0	+2.0
Effect of fatigue on reading time	-20.3	-35.6	—

(Table 4). This interaction can be expressed in two ways. The increase in reading time with the addition of the major fracture was far less when readers were tired (+13 s versus +41 s). Stated differently, the decrease in reading time when fatigued was far greater for the SOS condition (-42 s) than the non-SOS condition (-14 s).

There was a statistically significant interaction of fatigue (experiment) by the presence of a test fracture, [$F(1,38) = 20.63$, $P = 0.0001$] (Table 5). Once again, this interaction can be expressed in two ways. The increase in reading time when the test fracture was present was far less when readers were tired (+2 s versus +17 s). Stated differently, the decrease in reading time when tired was far greater for the SOS condition (-36 s) than the non-SOS condition (-20 s).

4 Discussion

In this research, we found that fatigue reduces detection accuracy for subtle fractures, whereas inclusion of multiple abnormalities reduces the range of reporting thresholds. Detection accuracy for obvious serious injuries was not affected by fatigue, but detection accuracy for subtle fractures was reduced. Considering both experiments with rested¹¹ and tired readers together, there was no statistically significant effect of the SOS manipulation on detection accuracy. There was also no statistically significant interaction of SOS treatment by fatigue with regard to detection accuracy.

There was a troubling inconsistency in the results for detection accuracy between analysis on individual experiments using the OR-DBM technique and the ANOVA, in which both experiments are included together. We found a significant SOS effect on detection accuracy for nonfatigued observers but not for fatigued observers in separate OR-DBM analyses, but no statistically significant interaction in the two-way ANOVA that included both experiments. The former combination of results from OR-DBM ought to be expressed in the ANOVA results as a significant interaction. The lack of a significant interaction may point to a lack of statistical power in testing interaction effects on detection accuracy. However, lack of power is not the only explanation for the apparent inconsistency.

In the previous nonfatigued SOS experiment, the OR-DBM procedure demonstrated a “meager at best” SOS effect on detection accuracy. AUC under non-SOS was 0.88 and 0.86 under SOS manipulation [P (one-tailed) = 0.09]. For sensitivity at specificity of 0.9, the difference was slightly more significant ($P = 0.07$). In any case, the SOS effect on detection accuracy was meager at best. In that experiment, the expectation from conventional wisdom is that serious injuries would produce more powerful SOS effects. We could not rule out that there was something there, but the SOS effects were only on the edge of significance. For the tired readers, there was no reduction in ROC area for detecting subtle test fractures when a major injury was present in the CT examination [AUC = 0.84 without major injury versus 0.83 with major injury; P (one-tailed) = 0.45]. The results for sensitivity at specificity of 0.9 were also not significant ($P = 0.23$). So for nontired readers, we got a meager on the edge of significance SOS effect while for tired readers we got no SOS effect using the individual OR-DBM analyses. When we combine the data into a single analysis, we get no main effect of SOS ($P = 0.37$). This is not surprising at all. What about the lack of interaction? The average SOS difference in the two experiments was in the same direction (nontired: non-SOS = 0.88 versus 0.86; tired: non-SOS =

0.84 versus 0.83). We could argue that neither is significant. So there is not actually much fuel for a significant interaction. At the same time, we do get a respectable main effect of fatigue ($P < 0.05$) with fatigued performance (AUC = 0.83) being lower than not fatigued (AUC = 0.88) overall. The apparent inconsistency is easily explained, and there is no real reason for concern about statistical power.

Looking at decision thresholds rather than accuracy reverses the pattern of results. There were no significant effects on the center of the FP ranges, but there were for the width of range of FPs associated with decision thresholds. While fatigue had no effect, the addition of a major injury reduced the range (width) of decision thresholds. The interaction of fatigue by SOS manipulation on the range was not statistically significant.

In all of these analyses on detection accuracy and decision thresholds, there are no statistically significant interactions of fatigue by SOS (including additional abnormalities). This suggests that the effects of fatigue and multiple abnormalities simply add to one another. Beyond that lack of statistical interactions, fatigue affects only detection accuracy whereas additional injury affects only the range of decision thresholds. In other words, the two factors affect different dependent variables. Both effects cause missed diagnoses, but the causes are different.

This straightforward picture does not hold when we turn to interpretation time. Decision thresholds and detection accuracy are the two major outcome variables that ROC methodology allows us to measure.^{22,23} For most studies of diagnostic performance in radiology, they are endpoints of analysis. Interpretation time is different. It can be an outcome in its own right, particularly where concern is with cost of image interpretation and workload distribution. On the other hand, interpretation time is also an intervening or mediating variable that indicates the extent of the perceptual and cognitive processes required for interpretation. Observed detection accuracy and decision thresholds must depend on the observed time needed for these underlying perception and cognition processes.

Our preliminary analysis of fatigue, which treated major injuries and subtle test fractures as a single dimension (no injury, subtle fracture, major injury, and both), indicated that fatigued readers do not increase reading time as the number of abnormalities present increases. Without fatigue, readers spent more time interpreting and reporting findings as the number of the injuries increased. However, when readers were fatigued, they did not. Without any fractures, reading time for not-fatigued and fatigued readers was about the same but differed with the addition of a subtle fracture. The difference increased with a major injury and increased further with both a major injury and a subtle fracture. In placing conditions along this dimension, we assumed that a major injury would have a greater impact than a subtle fracture. The ANOVA treated the presence of the major injury and the subtle fracture as orthogonal factors allowing their interaction with each other and with fatigue to be examined.

Fatigue reduced reading time by 28 s on average. The SOS manipulation of including a major injury increased reading time by 27 s, and the inclusion of a test fracture presence increased reading time by 10 s. In addition to these “main effects,” the ANOVA allowed the two-way interactions to be systematically interpreted (and those interpretations may qualify the interpretation of the main effects). The interaction of test fracture presence by major injury presence indicates that more time is needed

when both are present than would be expected from the increases in reading time associated with including just test fractures alone or just major injuries alone. The two interactions with fatigue (test fracture by fatigue and major injury by fatigue) indicate that fatigued readers do not take the extra time to deal with more injuries than rested readers do. Including a major fracture increases reading time by 41 s for rested readers but by only by 13 s for tired readers. Including a test fracture increases reading time by 17 s for rested readers but by only 2 s for tired readers. These interactions correspond to the flattening of reading time across examinations with more abnormalities for tired readers noted in the preliminary analysis.

It is interesting that while the endpoints of ROC analysis show no evidence of interactions of fatigue and multiple abnormalities, analysis of reading time, which reflects the perceptual and cognitive activities that support interpretation, showed marked interactions. Perhaps, we would think that interactions of experimental factors on interpretation time would be reflected in similar interactions on diagnostic accuracy and the decision thresholds, but they are not.

Let us take each part of the results in turn beginning with the SOS effect. Two types of SOS effects have been distinguished.⁹ Type I SOS effects seem to be the result of faulty pattern recognition.²⁴ Decreases in ROC accuracy occur with decreases in TP probability without changes in FP probability for each ROC point. Type I SOS is unrelated to changes in search behavior or inspection time.²⁵ Type II SOS effects are based, however, on reductions in visual search. ROC points move downward along the ROC curve owing to reductions in both TP rates and FP rates. The cause of unwillingness to report abnormality in type II SOS has been identified as reduced inspection. In this taxonomy, the SOS effect in the current experiment, more conservative reporting with reduced inspection time, is consistent with type II SOS effects.

This explanation that the SOS effect is due to shifts in decision thresholds as reflected by narrowing of the range of response applies to the current results with and without fatigue. (Although graphically, Fig. 3 suggests that this narrowing may be exacerbated with fatigue, this is not born out in a significant interaction of SOS by fatigue in the range of FPs associated by ROC points.) The same SOS effect occurs with and without fatigue, and the same reduction in search occurs with and without fatigued readers. If SOS effects on decision thresholds are unrelated to changes in inspection time, this may be consistent with the recent reports of SOS effects on decision thresholds in radiography, in which there are no changes in search time.^{7,9,10}

Under fatigue, the readers are holding the line on search time even though there are more abnormalities to find and report. More specifically, in the analysis of search time, the two interactions with fatigue—test fracture by fatigue and major injury by fatigue—indicate that fatigued readers do not take the extra time to deal with more injuries that rested readers do. For tired readers, the interactions correspond to the flattening of reading time even with more abnormalities. This is a new finding. Previous studies would not have found this difference when there was no more than one finding on an examination.^{5,6} In the study, in which there was more than one findings,⁷ no direct statistical comparison could be performed owing to differences in sampling of tired and nontired readers. They still get the major findings but miss some of the subtle ones. But fatigue has been shown time and again to reduce detection accuracy

for subtle abnormalities for residents (whose level of fatigue may exceed that of faculty radiologists).⁵⁻⁷

Previous studies of the effects of fatigue in radiology also measured inspection time.⁵⁻⁷ In two previous studies, total inspection time did not differ depending on fatigue status.^{5,6} Contrary to current findings, one previous study found significantly greater reading time for normal examinations than examinations with fractures (57 versus 47 s).⁵ In a recent study of tired readers in a classic SOS study in chest radiography,⁷ the addition of the distracting nodule did not affect inspection time, but the presence of a native abnormality required greater inspection time with 40 s required without native abnormalities and 55 s with native abnormalities. This did not differ from the results with rested readers.⁹ With rested readers, the presence of native abnormality increased reading time from 36 s without native abnormalities to 50 s with native abnormalities.

Refusal to increase search time with multiple abnormalities when tired may be related to an observation reported long ago by Christensen et al.²⁶ Christensen et al.²⁶ found that experts terminated search before the FP rate began to outstrip the TP rate whereas less experienced readers did not. Berbaum et al.⁸ referred to this phenomenon to explain the findings of Berbaum et al.²⁵ on the time course of SOS effects. They also found that once the rates for detecting actual abnormalities and for FP responses became equal, searches were often terminated. In other words, observers halt search before responses are as likely to be false as true. The reluctance is observed only when there are multiple abnormalities inspected by tired readers. Perhaps readers limit the increase in search time with more abnormalities as a countermeasure for this law of diminishing returns in radiology.

5 Conclusion

In this paper, we achieved something innovative in studying the statistical interaction of fatigue and multiple abnormalities. The SOS effect for both fatigued and not-fatigued readers with multiply injured patient CT images is similar—there is a shift in criteria (reduced TP and FP rates) rather than a change in performance. Fatigued performance was poorer than rested performance. Not surprisingly, overall performance increased in both conditions with more training, although readers at all levels of experience demonstrated the decision shift and reduced performance with fatigue. Though multiple abnormalities and fatigue appear to make independent contributions to observer performance, they interact in determining search time.

Clinically, if as our results demonstrate, fatigue negatively impacts search time and diagnostic accuracy, there is significant potential to impact patient care as lesions are likely to be missed as radiologists become tired. The fact that residents appear to be impacted more than experienced radiologists raises concerns even more perhaps, as they are more typically covering late hours and overnight shifts when fatigue is likely even more prevalent. It is clear that the issue of fatigue and educating radiologists about its impact on clinical work needs to become a national priority.

Disclosures

None of the authors have disclosures.

Acknowledgments

This work was supported by Grant No. R01 EB004987 from the National Institute of Biomedical Imaging and Bioengineering, Bethesda, Maryland.

References

1. M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction," *RadioGraphics* **35**, 1668–1676 (2015).
2. S. Rohatgi et al., "After-hours radiology: challenges and strategies for the radiologist," *Am. J. Roentgenol.* **205**, 956–961 (2015).
3. J. C. Hoffmann et al., "Combating the health risks of sedentary behavior in the contemporary radiology reading room," *Am. J. Roentgenol.* **206**, 1135–1140 (2016).
4. S. Waite et al., "Tired in the reading room: the influence of fatigue in radiology," *J. Am. Coll. Radiol.* **14**, 191–197 (2017).
5. E. A. Krupinski et al., "Long radiology workdays reduce detection and accommodation accuracy," *J. Am. Coll. Radiol.* **7**, 698–704 (2010).
6. E. A. Krupinski et al., "Do long radiology workdays affect nodule detection in dynamic CT interpretation?" *J. Am. Coll. Radiol.* **9**, 191–198 (2012).
7. E. A. Krupinski et al., "The impact of fatigue on satisfaction of search in chest radiography," *Acad. Radiol.* **24**(9), 1058–1063 (2017).
8. K. Berbaum et al., "Satisfaction of search in traditional radiographic imaging," in *The Handbook of Medical Image Perception and Techniques*, E. Samei and E. Krupinski, Eds., pp. 107–138, Cambridge University Press, New York (2010).
9. K. S. Berbaum et al., "Satisfaction of search in chest radiography 2015," *Acad. Radiol.* **22**, 1457–1465 (2015).
10. K. S. Berbaum et al., "The influence of a vocalized checklist on detection of multiple abnormalities in chest radiography," *Acad. Radiol.* **23**, 413–420 (2016).
11. K. M. Scharztz et al., "Trauma in CT: the role of severe injury on satisfaction of search revised," *J. Am. Coll. Radiol.* **13**, 973–978 (2016).
12. K. M. Scharztz et al., "WorkstationJ: workstation emulation software for medical image perception and technology evaluation research," *Proc. SPIE* **6515**, 65151I (2007).
13. J. T. Norbeck et al., "ACR-AAPM-SIIM technical standard for electronic practice of medical imaging," *J. Digital Imaging* **26**, 38–52 (2013).
14. B. J. Winer, *Statistical Principles in Experimental Design*, 2nd ed., pp. 309–684, McGraw-Hill Kogakusha, Tokyo (1971).
15. R. A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron* **1**, 3–32 (1921).
16. D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method," *Invest. Radiol.* **27**, 723–731 (1992).
17. S. L. Hillis, "Monte Carlo validation of the Dorfman–Berbaum–Metz method using normalized pseudovalues and less data-based model simplification," *Acad. Radiol.* **12**, 1534–1541 (2005).
18. OR-DB M-MRMC 2.5, University of Iowa Medical Image, 2006, <http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx> (22 September 2017).
19. D. D. Dorfman and K. S. Berbaum, "A contaminated binormal model for ROC data. II. A formal model," *Acad. Radiol.* **7**, 427–437 (2000).
20. BMDP3D, release: 8.0. Copyright 1993 by BMDP Statistical Software, Inc. Statistical Solutions Ltd., Cork, Ireland, <http://www.statsol.ie>.
21. BMDP2V, release v8.0; Copyright 1993 by BMDP Statistical Software, Inc. Statistical Solutions Ltd., Cork, Ireland, <http://www.statsol.ie>.
22. E. J. Potchen, "Measuring observer performance in chest radiology: some experiences," *J. Am. Coll. Radiol.* **3**, 423–432 (2006).
23. N. A. Obuchowski, "Receiver operating characteristic curves and their use in radiology," *Radiology* **229**, 3–8 (2003).
24. K. S. Berbaum et al., "Satisfaction of search from detection of pulmonary nodules in computed tomography of the chest," *Acad. Radiol.* **20**, 194–201 (2013).
25. K. S. Berbaum et al., "Time course of satisfaction of search," *Invest. Radiol.* **26**, 640–648 (1991).

26. E. E. Christensen et al., "The effect of search time on perception," *Radiology* **138**, 361–365 (1981).

Elizabeth A. Krupinski, PhD, is a professor and vice chair of radiology in the Department of Radiology and Imaging Sciences, Emory University. Her interests are in medical image perception, observer performance assessment and human factors as they pertain to radiology, pathology, and telemedicine applications.

Kevin Schartz, PhD, is a research associate professor in the Department of Radiology, University of Iowa. His educational background includes a PhD in experimental psychology specializing in psychophysics and a MCS degree specializing in software engineering. He is author of the WorkstationJ software used for medical image perception and technology evaluation studies.

Mark S. Van Tassel, MD, is a radiologist at Radiology Montana. He did his undergraduate work at the University of UT, medical school at

the University of Iowa where he also did his residency and fellowship in neuroradiology.

Mark T. Madsen, PhD, is a professor at the University of Iowa in the Departments of Radiology and Physics & Astronomy. He is a medical physicist with experience in radiotracer kinetics, image processing, and image reconstruction.

Robert T. Caldwell, MFA, is retired from the University of Iowa, Department of Radiology with expertise in experimental design, maintaining the integrity of data collection, and critical reading, writing, and editing grant applications, reports, and scientific publications.

Kevin S. Berbaum, PhD, is a professor in the Department of Radiology, University of Iowa. He is an expert in ROC research design and statistical analysis in observer-performance experimentation and the development of new ROC methods.