Impact of Bone Suppression Imaging on the Detection of Lung Nodules in Chest Radiographs: Analysis of Multiple Reading Sessions

S Schalekamp[a], B van Ginneken[a], CM Schaefer-Prokop[a,b], N Karssemeijer[a].

[a]Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands;
[b]Meander Medical Centre, Amersfoort, The Netherlands;

ABSTRACT

Observer studies are frequently performed to test new modalities. Correct study design is important to generate reliable results. Two most frequently used observer study designs are the sequential and the independent reading design. We investigated the effect of different observer study designs on reader performance results and statistical power. The study included multiple assessments of chest radiographs (CXR) with bone suppression images (BSI) for the detection of lung nodules. In a fully crossed study design 8 observers assessed first radiographs without and with BSI sequentially. Secondly they scored radiographs independently having BSI available from the beginning. Five months later, the same readers scored the same cases again in an independent reading session, completing the three scorings for CXRs with BSI. Observer performance was compared using multi reader multi case (MRMC) receiver operating characteristics (ROC). To estimate reader variance, Dorfman, Berbaum, Metz (DBM) variance component estimates were calculated. No significant difference between the sequential and the independent reading sessions could be found (p=0.51; p=0.61). Both reading designs showed increased performance with BSI, with a significant increase for the sequential and the independent reading session after five months (p=0.002; p=0.007). Total observer variance between sequential and independent reading design remained the same. A strong increase of uncorrelated components was found in the independent reading sessions, masking the ability to demonstrate differences in observer performance across modalities. In conclusion, results of the sequential and the independent study design did not show a significant difference. The independent study design had less power compared to the sequential study design due to a strong increase of uncorrelated variance components.

Keywords: Observer performance, study design, receiver operating characteristics, reader variability, bone suppression imaging, lung nodule, chest radiography.

1. INTRODUCTION

Before new imaging techniques can be implemented in the clinic, studies need to prove their added value. Observer studies are often used to investigate this added value. Since results can be influenced by the study design, it is important to choose a correct design. Two most frequently used observer study designs are the sequential design and the independent design. In a sequential design observers evaluate the new modality immediately after (sequentially) the current modality, providing two separate scores. In an independent design evaluation of the current and the new modality takes place in two separate reading sessions at different time points.

Some papers in the literature are favoring a sequential reading study design above an independent design(1, 2). Still the effect of sequential reading on observers' vigilance in detection performance studies remains unclear. Readers might increase their performance, trying to compete with the aided modality (for instance with computer aided detection). On the other hand, they might decrease their performance knowing that they can provide a second assessment with the aided modality. Also it is often questioned if the inevitably prolonged reading time in a sequential design improves the reader performance.

On the other hand, an independent reading study involves two separate reading sessions and images are evaluated twice. Readers may recall cases from the first evaluation, which could influence performance of the second evaluation. Randomization, counterbalancing and time between reading sessions are used to reduce this memory effect(3). The minimally required extent of this time lag is still unknown.

Observer studies are often performed, but only few studies have investigated the effect of study design on reader performance(1, 4). Mostly only either the sequential or the independent design is chosen. Only few studies incorporated both designs. These could not demonstrate a difference in effect size between a sequential design and an independent

Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment,
edited by Craig K. Abbey, Claudia R. Mello-Thoms, Proc. of SPIE Vol. 8673, 86730Y
© 2013 SPIE · CCC code: 1605-7422/13/$18 · doi: 10.1117/12.2002799

design [5, 6]. We conducted an observer study with PA and lateral chest radiographs (CXR), which included multiple reading modes to evaluate the added value of bone suppression imaging (BSI). The study consisted of a fully crossed design with a sequential and an independent reading design. In addition, after five months, all readers assessed the same cases in a reading session that consisted of an initial assessment of CXRs with BSI, sequentially followed by an assessment with help of computer aided detection marks, providing us with results of a third evaluation of CXR with BSI of the same data set. All the reading sessions incorporated the same readers, the same cases, and the same reviewing conditions. Using the reading data of these multiple assessments that applied both, sequential and independent reading design, in a counterbalanced and unbalanced way, we aimed to investigate the influence of study design on reader performance and statistical power.

2. METHODS

We performed an observer study with multiple reading sessions, involving 8 observers, assessing 111 chest radiographs with a solitary nodule and 189 controls, for the detection of lung nodules with and without bone suppression images.

Data

Three hundred cases were selected from four hospitals in The Netherlands. All images were obtained for clinical purposes. Posteroanterior and lateral radiographs of patients with a solitary nodule that was confirmed by a thoracic CT scan within 3 months time of the chest radiograph were included in the study. The conspicuity of the nodule on the PA radiograph was scored by an expert radiologist and a clinical researcher in consensus. Nodules needed to be visible on the PA radiographs (with knowledge of the CT). The size of the nodules was between 5 and 35 mm. Patients with multiple nodules, too obvious nodule or signs of other diseases than chronic pulmonary obstructive disease (COPD) were excluded. Age matched patients with a normal chest radiograph and a normal thoracic CT scan within 6 months were used as controls.

Software

Bone suppression images were generated using ClearRead BSI 2.4 (Riverain Technologies, Miamisburg, Ohio). This processing tool produces bone suppressed images that are identical in size and similar in gradation characteristics as the original chest radiograph. No special hardware or additional dose is needed to create the bone suppressed images.

Study Design

Five radiologists and three residents, from two institutions, participated in an observer study. The observer study included a sequential reading mode, where observers scored the original radiograph, immediately followed by a second scoring with the availability of BSI (sequential mode). Secondly, the same readers evaluated the same 300 cases, but now with BSI available from the beginning, providing a single score (independent mode 1). These two reading modes were balanced, meaning that all the observers evaluated in one session half of the cases in sequential mode and half of the cases in independent mode. In a later stage they reviewed, in a second reading session, the other half of the cases.

Five months later the same observers assessed the same cases again, scoring CXR with BSI independently (independent mode 2), followed by an evaluation with computer aided detection marks (results of that will not be used in this paper). This provided us with three assessments of the cases with BSI compared to an unaided reading (one sequential and two independent).

Observers were able to mark and score suspicious regions in the CXR using a continuous scale from 0 to 100. Before evaluating the cases, a training set of 40 cases was provided to get familiar with the review station, reading modes and the BSI. During this training observer received instant feedback from the researcher. None of the observer had previous experience with BSI. In between the reading sessions none of the observers received feedback, used BSI outside of the study or had insight into any study results.

Readings were carried out using a 30 inch 4K DICOM-calibrated LCD monitor (Flexscan SW3031W; Eizo, Ishikawa, Japan) in a darkened room, mimicking clinical reading conditions. The screen was large enough to review both PA and lateral radiograph side-by-side. We developed a case review system where processing tools were available, including zoom in/out, adjustment of window and level and grey scale inversion, and could be applied as warranted by the readers. The BSI was projected behind the original PA radiograph on the same monitor. The readers could toggle between the original and the BSI using a key on the keyboard, to easily review corresponding areas.

Statistics

Multi reader multi case (MRMC) receiver operating characteristics (ROC) was used for analysis. For that purpose the Dorfman, Berbaum and Metz method was used (DBM MRMC package v.2.33) [7-9]. Observer performance was measured by the area under the curve (AUC) for the readings without and with BSI. Significance of difference between reading without and with BSI was defined at $p < 0.05$. DBM provided variance component estimates for reader, case and treatment[7].

3. RESULTS

Observer Performance

Average area under the curve for the eight observers for the unaided reading was 0.855. With BSI, AUCs increased to 0.883 (p=0.002) and 0.874 (p=0.21), for the sequential reading and independent reading respectively. The average AUC for the independent reading session after five months was 0.887 (p=0.007) (Table 1). In the sequential reading all observer increased their performance. Four of the eight observers increased their performance in the first independent reading, compared to the unaided reading. In the reading session after 5 months again all observers performed better than in the unaided reading. No differences were found between the sequential and independent reading sessions (p=0.51 and p=0.65), as well as between the two independent readings (p=0.17).

	unaided	BSI sequential	BSI independent 1	BSI independent 2
average AUC	0.855	0.883	0.874	0.887
p		0.002	0.210	0.007

Table 1: The average area under the ROC curve in different readings (unaided; sequential; independent 1; independent 2).

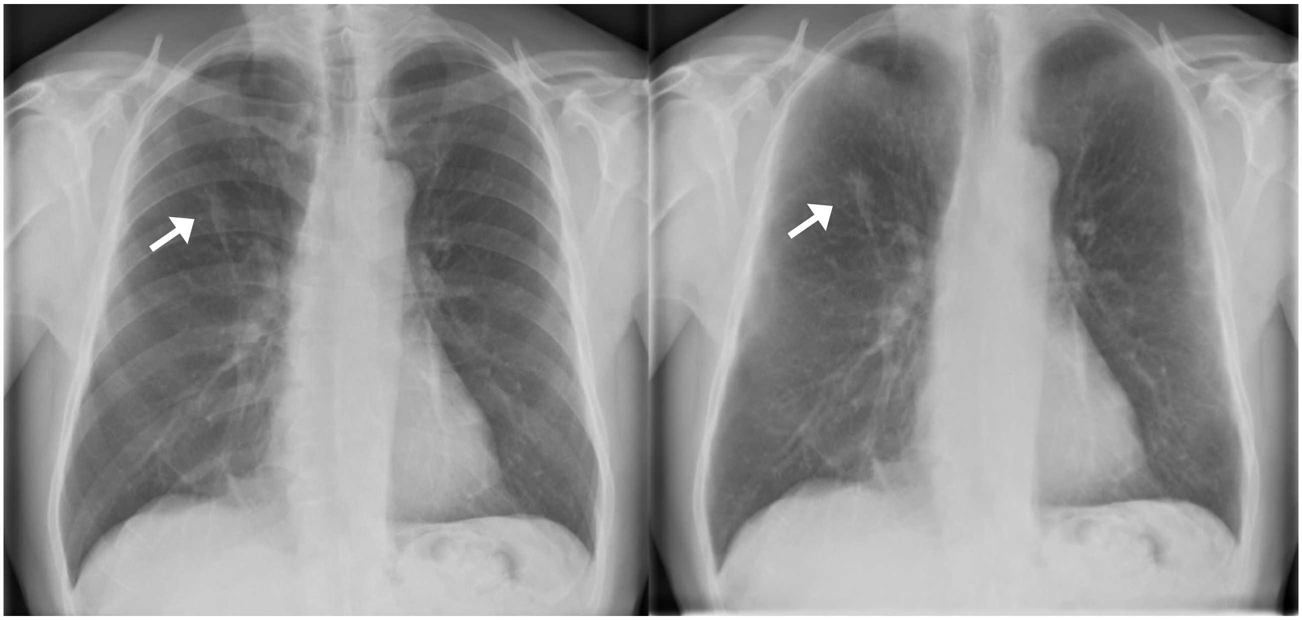

Figure 1: 65 year old male with a 27 mm spiculated lesion in the right upper lobe (white arrow). Without BSI (left) only three observers called the nodule suspicious with a score above 50. With BSI (right) seven of the eight observers noted the nodule suspicious with a score above 50, in the sequential reading. In the two independent reading sessions five and seven observers noted the nodule suspicious with a score above 50.

Analysis of Variance.

Comparing all the reading sessions with the unaided reading, ANOVA showed similar total variances for the different reading modes; 20.4×10^{-2} for the sequential reading, and 23.7×10^{-2} and 21.8×10^{-2} for the two independent reading sessions. The effect size of sequential reading was 0.028, for independent reading these values were 0.019 and 0.032. We found a shift from correlated to uncorrelated components in the independent readings, compared to the sequential reading. Correlated components in the sequential reading were 14.8×10^{-2} against 8.7×10^{-2} and 8.6×10^{-2} in the independent modes. Uncorrelated components increased from 5.5×10^{-2} in the sequential mode to 14.9×10^{-2} and 13.1×10^{-2} in the independent modes. All variance components are displayed in Table 2. Only the reader component (RC) component and the modality-by-reader-by-case (MRC) component contributed to this shift.

Variance Comp.	Sequential	Independent 1	Independent 2
correlated			
C	6.3 ×10^{-2}	6.9 ×10^{-2}	6.3 ×10^{-2}
R	1.3 ×10^{-3}	0.6 ×10^{-3}	1.3 ×10^{-3}
RC	8.4 ×10^{-2}	1.7 ×10^{-2}	2.2 ×10^{-2}
uncorrelated			
MC	5.3 ×10^{-3}	-0.4 ×10^{-3}	7.8 ×10^{-3}
MR	0.0 ×10^{-4}	2.5 ×10^{-4}	-1.5 ×10^{-4}
MRC	0.5 ×10^{-1}	1.5 ×10^{-1}	1.2 ×10^{-1}

Table 2: Variance components. C = case component; R = reader component; RC = reader-by-case component; MC = modality-by-case component; MR = modality-by-reader component; MRC = modality-by-reader-by-case component, including residual error. DBM generates unbiased variance components, which can be negative.

4. DISCUSSION

In this study we have consistently shown that BSI improves lung nodule detection performance for radiologists. We found a significantly increased detection performance with BSI in the sequential and independent reading mode 2. Although independent mode 1 yielded an increase in AUC, the difference with the unaided reading was not significant.

Several processes have been proposed that could influence observer performance results in an observer study with a sequential design. These include reader vigilance and interpretation time.

Reader vigilance indicates a potential effect on the observers' behavior because he knows that he has to interpret the image twice, unaided and aided. There may be two reasons for a change in behavior affecting the performance in the unaided baseline condition. The first is that the observer might perform worse knowing, that there will be a second chance to detect an abnormality. Secondly, the observer could also become more vigilant and increase his performance because he is trying to compete with the aided modality. Whether these effects play a role, however, cannot be determined from our study design. To investigate reader vigilance effect, a study should include an unaided reading in a sequential mode and an unaided reading in an independent mode. Another option would be to randomly show a sequential aided reading or not. In this design the observer will not know whether there is a sequential reading. In our study both independent reading sessions with BSI were followed by a second sequential assessment with CAD, and therefore could be affected by reader vigilance.

When studies are conducted using the sequential reading mode it is often argued that readers have an advantage in the aided mode just because they have more time to evaluate exams. In our study, reader may have reported more abnormalities with BSI, not as an effect of BSI but because of more reading time spent. Though research has shown that no correct false negative decisions for the detection of lung nodules in chest radiographs are made if interpretation of an area takes more than 3 seconds (10), still it can be argued that when they have to interpret the image again in the aided reading, the whole interpretation process starts over, resulting in a "second look" which might be different than prolonged reading time. In our study the total reading time for the sequential reading was on average indeed longer than the reading time for the independent readings (192 minutes versus 137 minutes). But we found similar results for the reading with BSI in a sequential mode and an independent mode. This suggests that in our study observers did not benefit from having a "second look", nor do the results show detrimental effects from having a prolonged reading time.

Another effect that possibly can influence observer study performance results is a learning effect. Because of extensive use of BSI, observers learn how to use the technique more optimally, gradually leading to improved performance. The steepness of the learning curve is different for each task, and unknown for bone suppression imaging. Since performance was best for the reading session after 5 months, a learning effect might indeed have played a role in our study.

Recall of cases by the observers might bias observer performance, when cases are evaluated at two different time points. Lengthening the time between the reading sessions might reduce this memory effect. But it is not known how long this time lag should be. Also it is imaginable that this memory effect is dependent on the difficulty and length of the task. For instance, readers may recall images more easily when they read a small set of images. By counterbalancing a study the memory effect should be the same for the tested modalities. In our study we used counterbalancing for the sequential reading and the independent mode 1. It can be argued that remembrance of cases from the independent reading only can influence the results for the sequential reading in a positive manner. This does not hold for the other way around; the independent reading session could not be positively influenced by the recall of cases from the sequential reading, since the same information was provided in the independent reading.

The observers also evaluated the same cases after a time period of five months. We believe that this time period was long enough to exclude any possible effects of memorizing cases. Although observers did not receive any information about study results or reviewed study cases outside of the observer study, they did perform best in the reading session after 5 months, though difference were not statistically significant. Recall of cases is unlikely to be responsible for this finding.

An important drawback in an independent reading design is reader variability. When evaluating the same data at two different time points, almost certainly different findings will be demonstrated. This effect is not only seen between observers (interobserver variability), but also within the same observer (intraobserver variability). In our study there were cases were an observer marked a suspicious lesion with a score of 100, while overlooking the same lesion in the independent reading. Figure 2 demonstrates a large variation in scores comparing the independent scorings with the unaided scoring. Even though this variation is large, also between the two independent readings (Figure 2), overall performance remained roughly the same. We found no difference in performance between the independent reading and the sequential reading (p=0.51; p=0.65). This confirms results of previous studies that compared independent reading with sequential reading results. Neither of those studies could demonstrate a significant bias (1, 4-6).

Although performance might not be affected, statistical power of experiment can suffer from this variation. In MRMC studies ANOVA analysis is used to estimate the variance in the study. This variance can be split up into variance components. The DBM method distinguishes six variance components. There is a pure case component, a pure reader component, a case x reader component; the so called correlated components. And there is a modality x case component, a modality x reader component, a modality x reader x case component; the so called uncorrelated components.

Correlated components are components that do not contribute to uncertainty in the measurements of differences in performance across modalities. Uncorrelated components do have influence on the uncertainty in the measurements of difference in performance across modalities. When we split the variance into DBM variance components, it is interesting that the total reader variability for the different study designs remained roughly the same. This was also found to be the case in a study by Beiden et al 2002. Although total variability remained the same, we found an increase in correlated components for the sequential reading, and an increase in uncorrelated components for the independent readings. Because of the increase in uncorrelated component for the independent reading, the uncertainty of the measurements increased, resulting in a loss of power. The sequential reading was not affected and is therefore the more powerful study design in MRMC studies.

5. CONCLUSION

We have shown that observer performance for the detection of lung nodules consistently improves with use of bone suppression imaging. Increase of performance for reading cases sequentially or independently are similar. In a second unbalanced independent reading session after five months, results are still comparable. All in all, a sequential study design is preferred over an independent design. The sequential study design is more practical, reduces total evaluation time, and is less affected by intra-reader variability. Both designs showed similar effect sizes, with more statistical power for the sequential design.

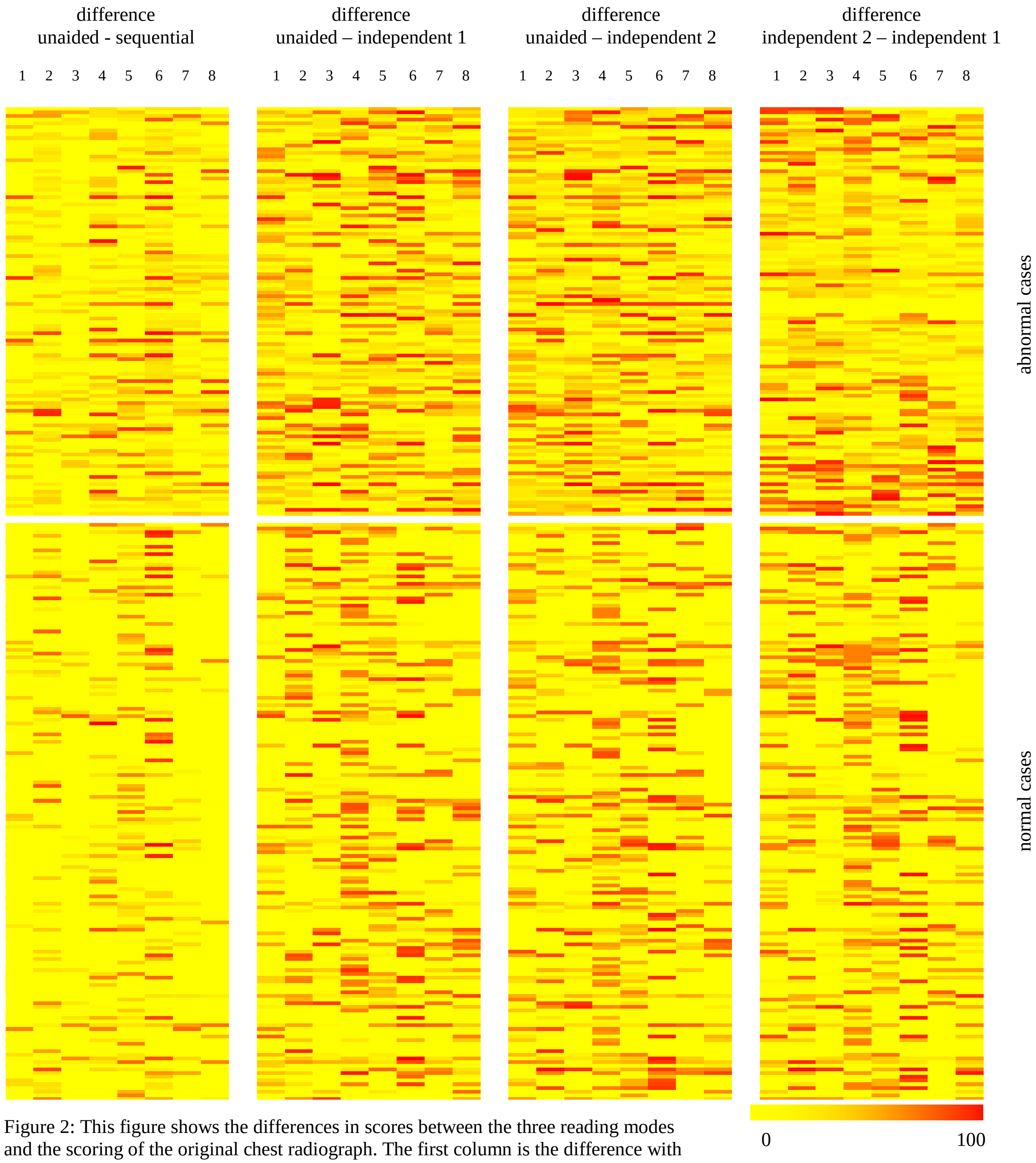

Figure 2: This figure shows the differences in scores between the three reading modes and the scoring of the original chest radiograph. The first column is the difference with the sequential mode, the second column is the difference with the independent mode 1, and the third column is the difference with the independent mode 2. On the Y-axis, on each line the individual cases are plotted. The cases with a nodule (top) are separated from the normal cases (bottom) with a white bar. Each difference diagram consists of eight columns, representing the eight observers. Yellow cells depict no difference between the reading modes for that particular case in that particular reader, red cells mark the greatest possible difference between the two reading (100 points in suspiciousness score). The sequential reading session shows remarkably less differences with the unaided reading, compared to the two independent reading sessions. The difference diagram of the independent readings (fourth column) is similar as the difference diagram of the unaided and one of the independent readings (second and third column), meaning that there is also a large variance between the two independent reading sessions.

6. REFERENCES

[1] Beiden, S. V., Wagner, R. F., Doi, K., Nishikawa, R. M., Freedman, M., Lo, S. B., and Xu, X., "Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of components of variance," *Academic Radiology* **9**, 1036–43 (2002).

[2] Gallas, B. D., Chan, H.-P., D'Orsi, C. J., Dodd, L. E., Giger, M. L., Gur, D., Krupinski, E. A., Metz, C. E., Myers, K. J., Obuchowski, N. A., Sahiner, B., Toledano, A. Y., and Zuley, M. L., "Evaluating imaging and computer-aided detection and diagnosis devices at the fda," *Academic Radiology* **19**, 463–477 (2012).

[3] Metz, C. E., "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigative Radiology* **24**, 234–245 (1989).

[4] Obuchowski, N. A., Meziane, M., Dachman, A. H., Lieber, M. L., and Mazzone, P. J., "What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance?," *Academic Radiology* **17**, 761–767 (2010).

[5] Kobayashi, T., Xu, X.-W., MacMahon, H., Metz, C., and Doi, K., "Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs," *Radiology* **199**, 843–848 (1996).

[6] Hadjiiski, L., Chan, H.-P., Sahiner, B., Helvie, M. A., Roubidoux, M. A., Blane, C., Paramagul, C., Petrick, N., Bailey, J., Klein, K., Foster, M., Patterson, S., Adler, D., Nees, A., and Shen, J., "Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an roc study," *Radiology* **233**, 255–265 (2004).

[7] Dorfman, D. D., Berbaum, K. S., and Metz, C. E., "Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method," *Investigative Radiology* **27**, 723–731 (1992).

[8] Roe, C. A. and Metz, C. E., "Variance-component modeling in the analysis of receiver operating characteristic index estimates," *Academic Radiology* **4**, 587–600 (1997).

[9] Hillis, S. L., Berbaum, K. S., and Metz, C. E., "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Academic Radiology* **15**, 647–661 (2008).

[10] Manning, D., Barker-Mill, S. C., Donovan, T., and Crawford, T., "Time-dependent observer errors in pulmonary nodule detection," *British Journal of Radiology* **79**, 342–6 (2006).