Journal of
Medical Imaging

SPIEDigitalLibrary.org/jmi

Influence of study design in receiver operating characteristics studies: sequential versus independent reading

Steven Schalekamp
Bram van Ginneken
Cornelia M. Schaefer-Prokop
Nico Karssemeijer

SPIE

Influence of study design in receiver operating characteristics studies: sequential versus independent reading

Steven Schalekamp,[a,*] Bram van Ginneken,[a] Cornelia M. Schaefer-Prokop,[a,b] and Nico Karssemeijer[a]
[a]Radboud University, Nijmegen, Medical Center, Department of Radiology, Postbus 9101, 6500 HB Nijmegen, The Netherlands
[b]Meander Medical Center, Department of Radiology, Postbus 1502, 3800 BM Amersfoort, The Netherlands

Abstract. Observer studies to assess new image processing devices or computer-aided diagnosis techniques are often performed, but little is known about the effect of the study design on observer performance results. We investigated the effect of the sequential and independent reading design on observer study results with respect to reader performance and their statistical power. For this we performed an observer study for the detection of lung nodules with bone-suppressed images (BSIs) compared with original chest radiographs. In a fully crossed observer study, eight observers assessed a series of 300 radiographs four times, including one assessment of the original radiograph with sequential BSI and two independent reading sessions with BSI. Observer performance was compared using multireader multicase receiver operating characteristics. No significant difference between the effect of BSI in the sequential and the independent reading sessions could be found ($p = 0.09$; $p = 0.46$). Compared with the original radiographs, increased performance with BSI was significant in the sequential and one of the independent reading sessions ($p < 0.0001$; $p = 0.0007$), and nonsignificant in the other independent reading session ($p = 0.10$). A strong increase of uncorrelated variance components was found in the independent reading sessions, masking the ability to demonstrate differences in observer performance across modalities. Therefore, the sequential reading design is the preferred design because it is less burdensome and has more statistical power. [DOI: 10.1117/1.JMI.1.1.015501]

Keywords: observer study design; observer performance; receiver operating characteristics; reader variability; bone-suppressed images; chest radiography.

Paper 13026PR received Dec. 16, 2013; revised manuscript received Jan. 28, 2014; accepted for publication Jan. 29, 2014; published online Apr. 23, 2014.

1 Introduction

Observer studies are often used to compare two modalities, or to determine the effect of a reading aid such as an image processing device or a computer-aided diagnosis (CAD) technique. The two evaluations of the new and current modality in an observer study are often referred to as reading mode 1 and reading mode 2. The two most frequently used observer study designs are the sequential reading design and the independent reading design. In a sequential reading design, reading mode 2 is evaluated immediately after (sequentially) reading mode 1. Thus, the reader provides two assessments for each case before moving on to the next case. In an independent reading design, the evaluation of the two reading modes takes place in two separate reading sessions at different time points. Both designs potentially suffer from bias.

In a sequential reading design, observers' vigilance may affect detection performance. Readers might increase their performance in the unaided modality, trying to compete with the aided modality (for instance with CAD). On the other hand, they might decrease their performance in the unaided mode, knowing that they have a second chance to provide a correct assessment with the aided modality. Also, it is often questioned if the inevitably prolonged reading time per case in a sequential design improves the reader performance (i.e., one unaided interpretation followed by an aided interpretation).

As mentioned above, an independent reading study involves two separate reading sessions at different time points. Therefore, readers may recall cases from the first evaluation, which could influence performance of the second evaluation. Randomization, counterbalancing, and a certain time interval between reading sessions are used to reduce this memory effect.[1] The minimally required extent of the time lag between two reading sessions is still unknown.

Reader variability is another important factor that should be dealt with in a study design.[2,3] There exists a wide variability not only between observers (interreader variability) but also within observers (intrareader variability). Major factors contributing to interreader variability are experience, visual skill, fatigue, motivation, and the fact that some readers are more aggressive than others in their decisions.[4,5] Receiver operating characteristics (ROC) analysis takes the effect of variable decision thresholds into account by calculating the sensitivity at various levels of specificity.[6] To account for interreader variability because of other factors, multireader multiplecase (MRMC) studies should be performed.[7] With the fully crossed MRMC design, where every reader reads every case in each modality, variability in

*Address all correspondence to: Steven Schalekamp, E-mail: steven.schalekamp@gmail.com

the study can be measured, and this information can be used to extrapolate the results to a wide range of readers and other case samples. Dedicated software, which accounts for the variability of readers, variability of cases, and the correlation of reader scores within and across the modalities, should therefore be used for analysis.

An important drawback of an independent reading design is that it is more prone to intrareader variability, because readers may judge a case very differently when it is presented for a second time in a new session. A sequential reading design does not have this drawback, because it builds in a correlation between the unaided and the aided scores.

Observer studies are often performed, but only few studies have investigated the effect of study design on reader performance.[8,9] Normally, either the sequential or the independent design is chosen. Only few studies have incorporated both designs. These could not demonstrate a difference in effect size between a sequential design and an independent design.[10,11] Therefore, some papers in the literature are favoring a sequential reading study design over an independent design.[7,8] Because of their efficiency, sequential studies are much less time-consuming and easier to conduct. However, evidence that both study designs lead to the same outcomes is still weak. Therefore, the purpose of this study was to provide more experimental data to compare the two study designs.

We conducted an observer study which included multiple reading modes to assess the effect of bone-suppressed images (BSIs) on the detection of lung nodules in chest radiographs (CXRs). The study consisted of a fully crossed design with a sequential and an independent reading design. In addition, after 5 months, all readers assessed the same cases in a third reading session that consisted of an initial assessment of CXRs supplemented by BSI, sequentially followed by an assessment with availability of CAD marks. Thus, in total we were able to compare three evaluations of CXR with the availability of BSI to an unaided reading. All the reading sessions involved the same readers and the same cases and were obtained under 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 determined the influence of study design on the measured effect of BSI and statistical power.

2 Methods

The observer studies involved eight observers of varying experience. The study data consisted of 111 chest radiographs with a solitary nodule and 189 controls. For both diseased and control cases, computed tomography (CT) provided the reference standard regarding the presence of nodules. The observer studies included one sequential reading design and two independent reading designs. Data from the reading study with the sequential reading design have been previously published.[12] In this article, we compare the previously reported results with two new independent readings of the same set of cases by the same readers. This study uses nonparametric ROC analysis as opposed to previously reported results, since we use different software for analysis which does not allow for parametric analysis of the data. [Our study results will be made publicly available through the iMRMC software (code.google.com/p/imrmrc/). This software package from the Food and Drug Administration (FDA) can be used to analyze data from ROC studies].

2.1 *Data*

Three hundred cases were selected from four hospitals in The Netherlands. All images were obtained for clinical purposes. Posteroanterior (PA) and lateral radiographs of patients with a solitary nodule whose presence was confirmed by a thoracic CT scan within 3 months of the chest radiograph were included in the study. One hundred eleven chest radiographs contained a single nodule; the other 189 radiographs functioned as controls. The conspicuity of the nodule on the PA radiograph was scored by an expert radiologist (>15 years of experience) 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 ranged from 5 to 35 mm. Patients with multiple nodules, too obvious nodules, or signs of diseases other than chronic pulmonary obstructive disease were excluded. Radiographs of patients over 40 years old with a normal thoracic CT scan within 6 months of the radiographs were used as controls. The total study group consisted of 300 cases.

2.2 *Software*

The BSIs were generated using ClearRead BSI 2.4 (Riverain Technologies, Miamisburg, Ohio). This processing tool produces BSIs that are identical in size and similar in gradation characteristics to the original chest radiograph. No special hardware or additional dose is needed to create the BSIs. The software has US FDA approval.

2.3 *Study Design*

Five radiologists and three residents, from two institutions, participated in the observer study. The first part of the observer study included a sequential reading mode: observers scored the original radiograph (unaided mode), immediately followed by a second scoring with the availability of BSI (sequential mode). Second, 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. After a minimum of 1 week 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 the latter will not be used in this article). Thus, the complete study provided us with three assessments of the cases with use of BSI and one unaided assessment using CXR alone (Fig. 1).

Observers were able to mark and score suspicious regions in the CXR using a continuous scale from 0 to 100. Scores of suspiciousness and localization were recorded digitally. 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 training, observers received instant feedback from the researcher. None of the observers had previous experience with BSI. In between the reading sessions, none of the observers received feedback, nor did they use BSI outside of the study or have insight into any study results.

Readings were carried out using a 30-inch 4K DICOM (Digital Imaging and Communications in Medicine) calibrated LCD monitor (Flexscan SX3031W; Eizo, Ishikawa, Japan) in a

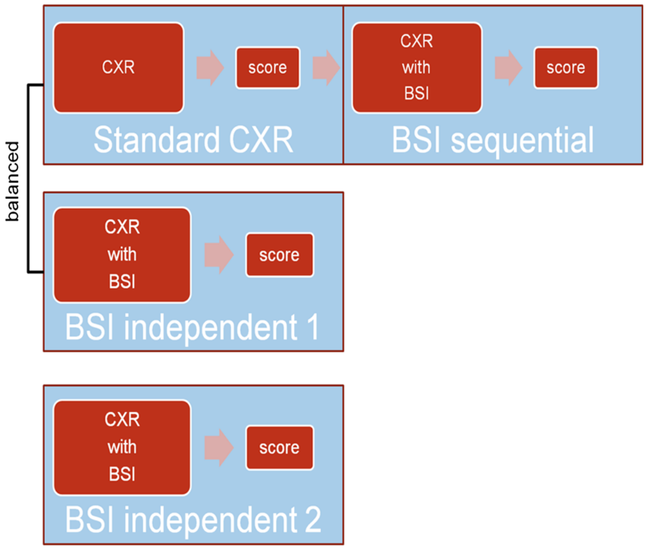

Fig. 1 Reading modes. The chest radiographs were scored four times by the observers. The first observer study is a sequential design where the radiographs were scored without bone-suppressed image (BSI) and with BSI within one reading session. Further there were two independent scorings with BSI; one that was balanced with the sequential reading design (BSI independent 1), and one after a 5-month period (BSI independent 2).

darkened room, mimicking clinical reading conditions. The screen was large enough to review both the PA and lateral radiograph side-by-side. For display, we used a workstation developed in-house. Processing tools were available, including zoom in/out, adjustment of window and level settings, and gray scale inversion. These tools could be applied without restriction. 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.

2.4 *Statistical Analysis*

2.4.1 *MRMC ROC analysis*

The MRMC ROC analysis was used for statistical analysis. For analysis, we used the iMRMC software package (code.google.com/p/imrmrc/; version 2.0b), which is developed at the FDA.[13,14] The software uses a nonparametric method to estimate the area under the ROC curve and calculate p-values.[15] Observer performance was measured by the area under the ROC curve for the readings without and with BSI. Significance of differences between reading without and with BSI was defined at $p < 0.05$. The results of the MRMC analysis method are not limited to a single reader or single case set but are generalizable to a population of readers and a population of cases.

Variance in the study can be decomposed into different variance components. We chose to use variance components calculated by the Dorfman-Berbaum-Metz (DBM) method,[16] since this method is the most widely used and available in several publicly available software packages. The DBM method distinguishes six variance components. Three of these components are not dependent on the modality and are called correlated components. These three include a pure case component (C), a pure reader component (R), and a reader-by-case component (RC). The other three variance components depend on the modality and are called uncorrelated components. These include a modality-by-case component (MC), modality-by-reader component (MR), and a modality-by-reader-by-case component (MRC).

2.4.2 *Reading time evaluation*

Reading times per case were digitally recorded, counting from start of evaluation of the case until the saving of the scores. Median reading times were used for analysis of reading times per case, to filter out the effect of long reading, caused by interruptions of the reading session. Two minutes without any mouse movement was considered as idle time and removed from analysis of reading times. Also, total reading time of the study was calculated by multiplying the mean reading time per case with the total amount of cases.

2.4.3 *Agreement*

To quantify the correlation of the scores of the different study designs, we calculated the intraclass correlations (ICCs) between the readings without BSI and with BSI per observer and for the group of all eight observers. The ICC reflects the agreement between two measurements on a (semi) continuous scale. In our study, a perfect correlation (ICC = 1) would mean that the same scores per case were given for the unaided reading and the reading with BSI. To further explore the source of variability in measurements, we also calculated the ICCs for the normal and abnormal cases separately. SPSS software (version 20) was used for calculation of ICCs, using a two-way random model. Besides calculating the correlation between the scores without and with BSI within the same observer, we also calculated the correlation between different observers for the different reading sessions (interobserver agreement).

Selection of study images and study setup was waived by the institutional review board.

3 Results

3.1 *Observer Performance*

Average area under the curve (AUC) for the eight observers for the unaided reading was 0.827. With BSI, AUCs increased to 0.868 ($p < 0.0001$) and 0.847 ($p = 0.10$), for the sequential reading and independent reading respectively. The average AUC for the independent reading session after 5 months was 0.862 ($p = 0.0007$) (Table 1). No significant differences were seen between the sequential and independent readings with BSI ($p = 0.09$ and $p = 0.46$) or between the two independent readings with BSI ($p = 0.12$).

With sequential reading, all eight observers increased their performance. With the first independent reading, five of the eight observers increased their performance, compared with the unaided reading. In the second independent reading session after 5 months, again all eight observers performed better compared with the unaided reading.

3.2 *Reading Times*

Sequential reading lengthened the median reading time per case by 12 s from 22 to 34 s. Reading time of independent reading with BSI compared with unaided reading was virtually the same at 23 s. In the independent reading session after 5 months, reading time remained similar at 21 s per case. Individual reading times are displayed in Table 2.

Table 1 The average area under the ROC curve in different readings (unaided; sequential; independent 1; independent 2). *P* values were calculated with the Dorfman-Berbaum-Metz method; readings with BSI were compared with the unaided reading. SD = standard deviation.

	Unaided	Bone-suppressed image (BSI) sequential	BSI independent 1	BSI independent 2
Observer 1	0.794	0.848	0.836	0.826
Observer 2	0.850	0.881	0.851	0.871
Observer 3	0.884	0.902	0.871	0.898
Observer 4	0.764	0.824	0.830	0.821
Observer 5	0.839	0.900	0.837	0.877
Observer 6	0.847	0.900	0.841	0.892
Observer 7	0.792	0.811	0.841	0.853
Observer 8	0.844	0.880	0.869	0.858
Average (SD)	0.827(0.039)	0.868(0.036)	0.847(0.015)	0.862(0.028)
p		<0.0001	0.10	0.0007

Total reading time of the cases was on average 129 min (range 71 to 233) for the unaided reading session, with an extra of 63 min for the sequential BSI reading, resulting in a total of 192 min for the sequential reading design. The independent reading sessions took on average 146 and 128 min, respectively. The total reading time of the independent reading designs, therefore, accumulated to 275 and 257 min, respectively.

3.3 *Analysis of Variance*

Comparing all reading sessions with BSI to the unaided reading, analysis of variance (ANOVA) showed similar total variances for the different reading designs; 169×10^{-3} for the sequential reading design, and 180×10^{-3} and 172×10^{-3} for the two independent reading designs. The effect size for sequential reading was 0.041, and these values were 0.020 and 0.035 for the independent readings. We found a shift from correlated to uncorrelated components in the independent reading designs, compared with the sequential reading design. Correlated components in the sequential reading were 135×10^{-3} against 92×10^{-3} and 88×10^{-3} in the independent readings. Uncorrelated components increased from 34×10^{-3} in the sequential reading to 88×10^{-3} and 84×10^{-3} in the independent readings. Mainly the RC component and the modality-reader-case MRC component contributed to this shift. All variance components are displayed in Table 3.

Table 2 Median reading times per case in seconds. Reading times of BSI sequential mode include the reading times of the unaided reading.

	Unaided	BSI sequential	BSI independent 1	BSI independent 2
Observer 1	15	26	15	17
Observer 2	25	44	27	24
Observer 3	21	31	24	13
Observer 4	12	19	12	11
Observer 5	14	25	15	23
Observer 6	27	50	25	31
Observer 7	42	53	47	27
Observer 8	21	27	16	20
Average	22	34	23	21

Table 3 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. Unbiased variance components were generated, which can be negative.

Variance component	Sequential ($\times 10^{-3}$)	Independent 1 ($\times 10^{-3}$)	Independent 2 ($\times 10^{-3}$)
C	68	73	68
R	1.0	0.4	0.9
RC	66	19	19
Correlated	135	92	88
MC	4.9	0.8	7.0
MR	0.04	0.15	−0.09
MRC	29	87	77
Uncorrelated	34	88	84

3.4 *Agreement*

The ICC for the sequential reading design was 0.896. The ICCs for the independent reading designs were 0.719 and 0.715, respectively (Table 4). The independent readings mutually showed as much correlation as the independent reading with the unaided reading with an ICC of 0.717. The ICC for only normal cases was 0.779 for the sequential reading mode and 0.338 and 0.320 for the two independent reading modes. For abnormal cases, ICC was 0.871 for the sequential reading mode and 0.658 and 0.643 for the independent reading modes.

The ICC between readers was 0.633 for unaided reading. The interreader variability remained virtually the same for all the readings with BSI, with ICCs of 0.631, 0.632, and 0.666 for the sequential reading and the two independent readings, respectively.

4 Discussion

Results of our study confirm that BSI consistently improves lung nodule detection performance of radiologists in both the independent and the sequential reading design. We found a significantly increased detection performance with aid of BSI in the sequential reading mode and in one of the independent reading modes. In the other independent mode, an increase in AUC was also found, but in that case the difference with the unaided reading was not statistically significant.

A potential bias that could be introduced by the sequential design is the lengthening of the reading time per case. In our unaided reading, radiologists reviewed a radiograph for 22 s on average, which was prolonged by another 12 s per case in the sequential reading design. It is therefore conceivable that readers may have reported more abnormalities with BSI, not as an effect of BSI, but as an effect of lengthened interpretation of the image. However, other studies reported more false-positive decisions with increasing reading time,[17,18] and we did not find a significant performance difference when comparing the results of the independent and the sequential design, indicating that the prolonged reading time in the sequential design did not affect observer performance. This agrees with similar findings of a previous study that could not show performance differences for the detection performance of pulmonary nodules by limiting the reading time.[19]

Table 4 Intraclass correlation (ICC) per observer for each study design. The overall ICC is the ICC for all observers together.

	BSI sequential design	BSI independent design 1	BSI independent design 2
Observer 1	0.918	0.722	0.665
Observer 2	0.900	0.728	0.767
Observer 3	0.980	0.784	0.790
Observer 4	0.882	0.629	0.608
Observer 5	0.862	0.732	0.726
Observer 6	0.770	0.617	0.623
Observer 7	0.962	0.767	0.750
Observer 8	0.930	0.741	0.807
Overall	0.896	0.719	0.715

Although the reading time per case lengthens with the sequential design, the total reading time of sequential reading is shorter than of an independent reading design. In our study, the total reading time for the sequential reading design was on average 192 min compared with an average of 266 min for the independent reading design. The cause of longer total reading time for an independent reading design lies in the need for two distinct reading sessions.

Another effect that potentially influences observer study performance results is the learning effect. In our study, because of repetitive use of BSI, it is likely that observers learned over time how to use the technique more optimally, and therefore gradually improved performance. The steepness of the learning curve is different for each task, and is unknown for BSI. Since

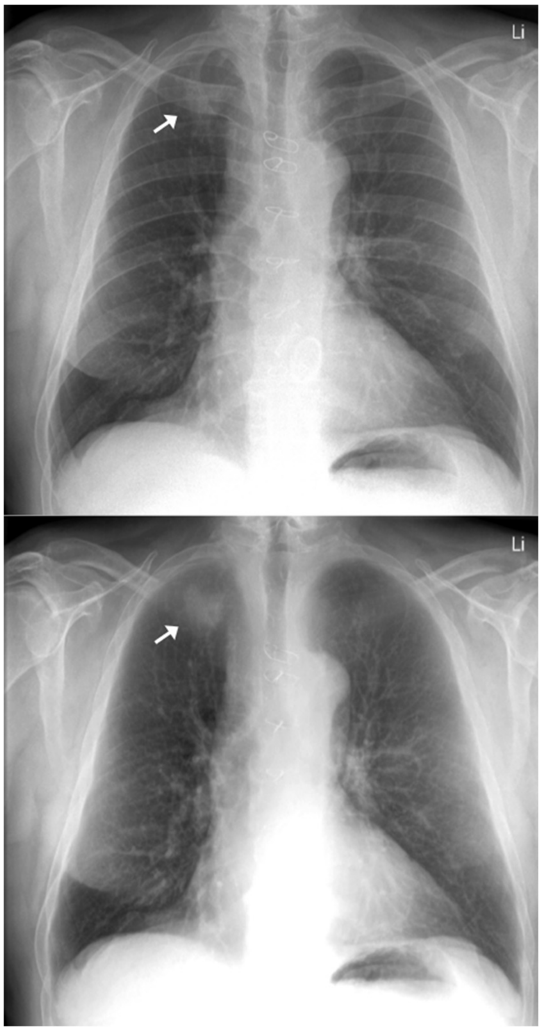

Fig. 2 A 63-year-old male with a 30-mm non–small cell carcinoma in the right upper lobe. A large variation between the observers and within observers was seen. Without BSI (upper) none of the observers called the nodule suspicious with a score above 50. With BSI (lower), five of the eight observers noted the nodule suspicious with a score above 50 in the sequential reading. In the two independent reading sessions, two and seven observers noted the nodule as suspicious with a score above 50. Only two observers noted the nodule in all three readings with BSI.

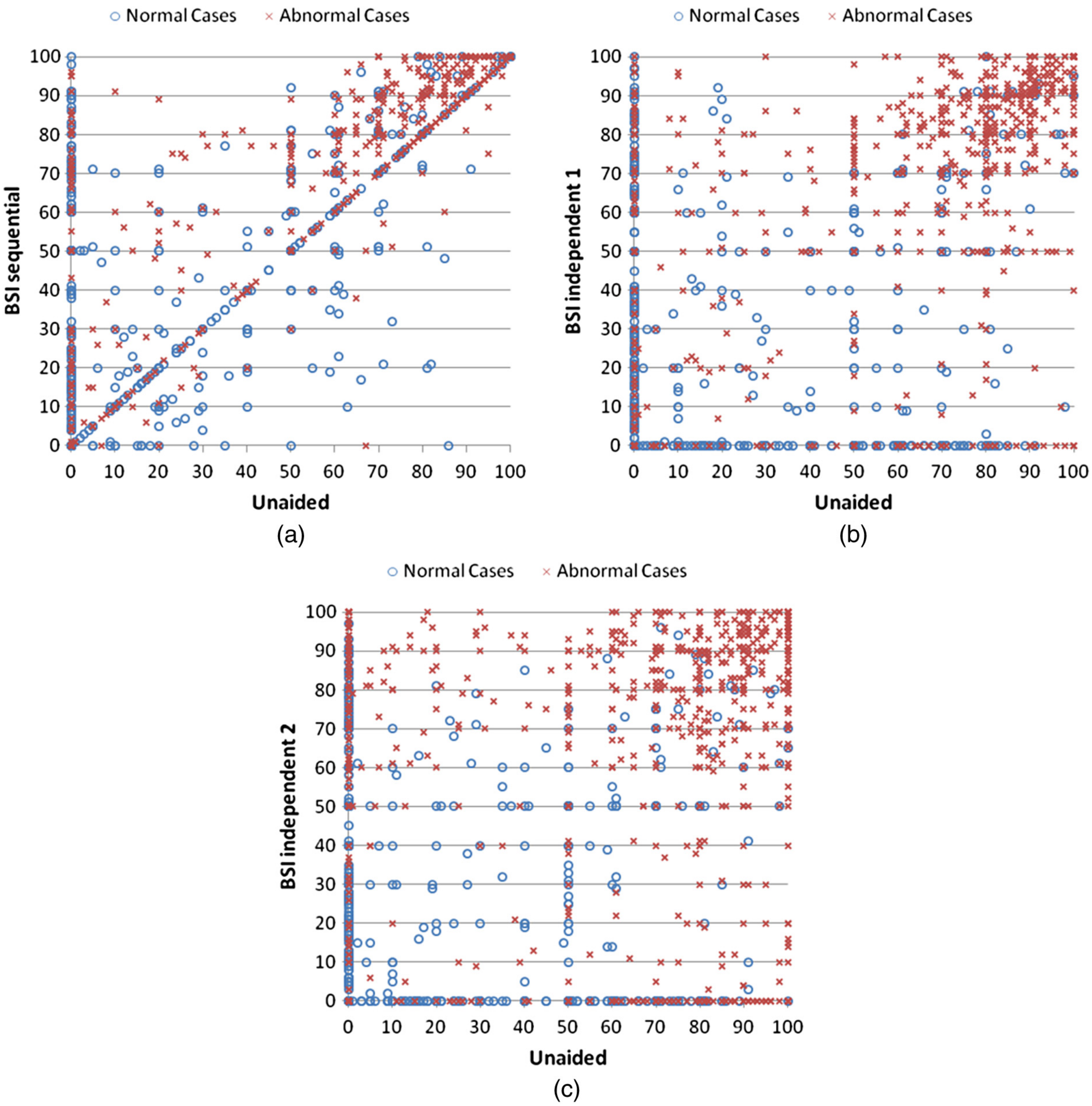

Fig. 3 These three charts show the correlation between the observer scores of the unaided reading with the scores of the readings with BSI. On the horizontal axis, the scores of the unaided reading are displayed (0 to 100). These scores are compared with the scores with BSI in the three study designs [one sequential scoring (A), two independent scorings (B and C)]. For the sequential design, a clear correlation between the scores without and with BSI can be observed.

performance was similar for the reading session after 5 months, it is unlikely that a learning effect has played a role in our study. It has to be noted that we have tried to minimize the learning by not providing feedback between the reading sessions. The observers also did not use the BSI software in clinical practice.

An important drawback in an independent reading design is its susceptibility for reader variability. When radiologists read the same data at two different time points, almost certainly different findings will be reported. This assumption is indeed confirmed by the data of our study (Fig. 2). This variability is seen not only between observers (interreader variability) but also within the same observer (intrareader variability). In our study, there were 14 cases where one observer had marked a suspicious lesion with a score of 100 with use of BSI, while overlooking the same lesion in another reading with BSI (Fig. 3). The ICCs demonstrated a smaller correlation of scores comparing the independent reading scores with the unaided scores. Interestingly, normal cases seem to contribute more to reader variation than abnormal cases. Even though the variation of scores in an independent design is quite large, also between the two independent readings, overall performance remained roughly the same. We found no significant performance difference between the independent readings and the sequential reading ($p = 0.09$; $p = 0.46$). Our results are in agreement with previously published findings that compared independent reading with sequential reading results. None of those studies could demonstrate a significant difference in effect size found in different study designs.[8–11]

Even though performance might not be significantly affected, statistical power is influenced by the variability of the readers. In MRMC ROC studies, ANOVA analysis is used to estimate the variance of the study. Decomposing the variance into DBM variance components, it is noteworthy that the total reader variability for the different study designs remained roughly the same. This has also been described in a study by Beiden et al.[8] 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 components for the independent reading, the uncertainty of the measurements increased, resulting in a loss of statistical power.

Finally, one other factor that could bias results in a sequential reading design is reader vigilance. Our study was not designed to assess whether reader vigilance indeed influences reader performance. To investigate the effect of reader vigilance, the study would need to include an unaided reading in a sequential mode and an unaided reading in an independent mode. Another option could be to randomly show a sequential aided reading or not. That way the observer would not know prospectively whether the assessment would be followed by a sequential assessment. Two previous studies which incorporated an unaided reading in a sequential and an independent reading design showed similar effect sizes.[10,20] This suggests that potential bias due to reader vigilance can be ignored.

In summary, we have shown that effect size measured in a study evaluating the detection of lung nodules is not influenced by the sequential or the independent design. In a second unbalanced independent reading session after 5 months, results were still comparable. Although an independent reading design is less susceptible to bias due to prolonged reading time and reader vigilance, up to now no significant bias could be demonstrated for a sequential design. The benefit of reduced intrareader variability leads to a preference for the sequential reading study design. Further on, the sequential study design is more practical because readers do not have to be invited twice, and it requires less evaluation time. Both designs showed similar effect sizes, with the advantage of higher statistical power for the sequential design.

References

1. C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Invest. Radiol.* **24**(3), 234–245 (1989).
2. A. Bankier et al., "Consensus interpretation in imaging research: is there a better way?," *Radiology* **257**(1), 14–17 (2010).
3. B. J. Hillman, "Acrin–lessons learned in conducting multi-center trials of imaging and cancer," *Cancer Imaging* **5**(A), S97–101 (2005).
4. J. G. Elmore et al., "Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy," *Radiology* **253**(3), 641–651 (2009).
5. C. A. Beam, P. M. Layde, and D. C. Sullivan, "Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample," *Arch. Intern. Med.* **156**(2), 209–213 (1996).
6. C. J. D'Orsi and J. A. Swets, "Variability in the interpretation of mammograms," *N. Engl. J. Med.* **332**, 1172–1173 (1995).
7. B. D. Gallas et al., "Evaluating imaging and computer-aided detection and diagnosis devices at the fda," *Acad. Radiol.* **19**(4), 463–477 (2012).
8. S. V. Beiden et al., "Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of components of variance," *Acad. Radiol.* **9**(9), 1036–1043 (2002).
9. N. A. Obuchowski et al., "What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance?," *Acad. Radiol.* **17**(6), 761–767 (2010).
10. T. Kobayashi et al., "Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs," *Radiology* **199**(3), 843–848 (1996).
11. L. Hadjiiski et al., "Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an roc study," *Radiology* **233**(1), 255–265 (2004).
12. S. Schalekamp et al., "Bone suppressed images improve radiologists' detection performance for pulmonary nodules in chest radiographs," *Eur. J. Radiol.* **82**(12), 2399–2405 (2013).
13. B. D. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.* **13**(3), 353–362 (2006).
14. B. D. Gallas et al., "A framework for random-effects roc analysis: biases with the bootstrap and other variance estimators," *Commun. Stat. A-Theory* **38**(15), 2586–2603 (2009).
15. S. L. Hillis, K. S. Berbaum, and C. E. Metz, "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Acad. Radiol.* **15**(5), 647–661 (2008).
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**(9), 723–731 (1992).
17. D. Manning et al., "Time-dependent observer errors in pulmonary nodule detection," *Br. J. Radiol.* **79**(940), 342–346 (2006).
18. R. S. Saunders and E. Samei, "Improving mammographic decision accuracy by incorporating observer ratings with interpretation time," *Br. J. Radiol.* **79**(Special Issue 2), S117–S122 (2006).
19. F. Li et al., "Small lung cancers: improved detection by use of bone suppression imaging–comparison with dual-energy subtraction chest radiography," *Radiology* **261**(3), 937–949 (2011).
20. Deus Technologies. Rapidscreen rs-2000; FDA approval study, http://www.accessdata.fda.gov/cdrh_docs/pdf/P000041b.pdf (2001).

Steven Schalekamp is a PhD candidate at the Radboud University Medical Center in Nijmegen, The Netherlands. He received his medical degree from the Free University in Amsterdam in 2011. As part of the Diagnostic Image Analysis Group (www.diagnijmegen.nl), he is currently investigating the effect of advanced image processing methods and computer-aided detection of lung nodules on observer performance in chest radiography.

Biographies of the other authors are not available.