HHS Public Access

Author manuscript

Acad Radiol. Author manuscript; available in PMC 2021 February 01.

Published in final edited form as:

Acad Radiol. 2020 February ; 27(2): 244–252. doi:10.1016/j.acra.2019.03.014.

Evaluation of pseudo-reader study designs to estimate observer performance results as an alternative to fully crossed, multi-reader, multi-case studies

Rickey E. Carter, PhD[a], **David R. Holmes III, PhD**[b], **Joel G. Fletcher, MD**[c], **Cynthia H. McCollough, PhD**[c]

[a]Department of Health Sciences Research, Mayo Clinic, 4500 San Pablo Road South Jacksonville, FL 32224

[b]Department of Physiology and Biomedical Engineering, Mayo Clinic, 200 First Street SW Rochester MN 55905

[c]Department of Radiology, Mayo Clinic, 200 First Street SW Rochester MN 55905

Abstract

Rationale and Objectives: To examine the ability of a pseudo-reader study design to estimate the observer performance obtained using a traditional fully crossed, multi-reader, multi-case (MRMC) study.

Materials and Methods: A ten-reader MRMC study with 20 computed tomography datasets was designed to measure observer performance on four novel noise reduction methods. This study served as the foundation for the empirical evaluation of three different pseudo-reader designs, each of which used a similar bootstrap approach for generating 2000 realizations from the fully crossed study. Our three approaches to generating a pseudo-reader varied in the degree to which reader performance was matched and integrated into the pseudo-reader design. One randomly selected simulation was selected as a “mock study” to represent a hypothetical, prospective implementation of the design.

Results: Using the traditional fully crossed design, figures of merit (FOM) (95% CIs) for the four noise reductions methods were 68.2 (55.5 – 81.0), 69.6 (58.4 – 80.8), 70.8 (60.2 – 81.4), and 70.9 (60.4 – 81.3), respectively. When radiologists’ performances on the fourth noise reduction method were used to pair readers during the mock study, there was strong agreement in the estimated FOMs with estimates using the pseudo-reader design being within +/−3% of the fully crossed design.

*Corresponding author 4500 San Pablo Road South, Jacksonville, FL 32224, Telephone: 904-953-0381, carter.rickey@mayo.edu.

Declaration of Interest: Dr. McCollough received industry funding support from Siemens Healthcare, unrelated to this work. The other authors have nothing to declare.

Conclusion: Fully crossed MRMC studies require significant investment in resources and time, often resulting in delayed implementation or minimal human testing before dissemination. The pseudo-reader approach accelerates study conduct by combining readers judiciously and was found to provide comparable results to the traditional fully crossed design by making strong assumptions about exchangeability of the readers.

Keywords

multi-reader; multi-case study; observer performance; study design

Introduction

With the ongoing development of new medical image acquisition and reconstruction methods comes the need to objectively measure how human (e.g., radiologist) performance changes with the new or altered appearance to medical images (1, 2). This is an essential step in confirming the diagnostic accuracy of the imaging technique prior to its adoption in clinical practice. To date, the standard approach is to utilize a multi-reader, multi-case (MRMC) study design wherein a large number of readers evaluate a large number of cases examining different imaging alternatives. MRMC designs are categorized as fully crossed when each reader reviews every patient case using each imaging strategy. The resources to conduct such a study are exhaustive. For example, if there are 10 radiologists reviewing 100 cases for five different imaging strategies (e.g., radiation dose levels), each radiologist must review 500 datasets, with the total reading burden for the study being 5000 total reader interpretations. Frequently, these reading interpretations need to be spread out over time, sometimes months apart, in order to minimize bias related to recall of conspicuous attributes in the patient/dataset, so fully crossed MRMC studies do not provide an expeditious path to evaluating new imaging alternatives. The evidence base for new imaging strategies may lag considerably behind technology development., and in some cases,the technologies under study may no longer considered state-of-the-art by the time the results of MRMC studies are disseminated.

To address these limitations, alternative designs to help accelerate the conduct of the human observer performance studies have been proposed(3–5). A pseudo-reader study design is an emerging alternative to a fully crossed MRMC study based on the concept of a Latin square experimental design(3). This design distributes the reading assignments over a large number of readers using an a priori plan to combine individual readers back into one or more virtual, or as now denoted “pseudo-“, reader(s) for analysis. While the design is hypothesized to allow for much greater flexibility with reading schedules and accelerate the time required to quantify human observer performance, the operating characteristics, limitations, requirements and performance of this approach are not fully described or understood.

In 2016, our team, in collaboration with the American Association of Physicists in Medicine (AAPM) and the National Institute of Biomedical Imaging and Bioengineering, conducted the 2016 Low Dose CT Grand Challenge.(3) In this challenge, multiple institutions received low-dose CT images or projection data from contrast-enhanced abdominal CT examinations, and returned denoised images, with the winner to be declared based on the correct visual

detection of hepatic metastases. Because of the relatively short time frame over which the challenge was conducted, a pseudo-reader approach was used to assess the submitted image denoising and iterative reconstruction approaches. The purpose of the present study is to empirically determine the level of agreement between the pseudo-reader design and a fully crossed MRMC study design by conducting a follow-up study on selected submissions to the AAPM CT Grand Challenge.

Methods

Institutional review board approval was obtained for this HIPAA-compliant study. Radiologists participating in this study provided oral consent according to the instructions related to the institutional review board approval.

Validation Study Design

We selected images returned by four of the sites participating in the AAPM CT Low Dose Grand Challenge for inclusion in this study; 2 sites performed projection space iterative reconstruction to reduce noise and 2 sites used image space noise reduction techniques. This reuse of the AAPM CT Low Dose Grand Challenge was consistent with the signed data sharing agreement associated with the challenge(6). To blind this study to the original competition results, the noise reduction method (NRM) used by each of the sites is simply referred to by the designation NRM A – D (Figure 1). As part of the original grand challenge, these four NRMs were applied to 20 simulated low-dose CT patient datasets, which had been prepared by using a validated technique to insert noise into the measured projection data(7). All radiologist reader interpretations reported herein are unique to the current work, and were not reported as results in the Grand Challenge(3). Details relating to reference standards for the contrast-enhanced CT data in the study has been previously published(3).

A total of 10 radiology trainees (senior residents and fellows) volunteered for participation as readers in this study and provided oral informed consent prior to participating. Prior to initiation of reads for the study, the readers received standardized training on participation in MRMC studies by a radiologist co-investigator (J. G. F. with experience in reader training(8, 9) and 20 years as staff abdominal radiologist), including training on the correct use of confidence scales and the operation of the custom reader interpretation workstation. The primary task for this study was the detection of an undisclosed number of hepatic metastases spread over the 20 patients. The MRMC design utilized the standard fully crossed design such that each reader would interpret a total of 80 datasets. Four reading sessions were required for each radiologist using randomized reading worklists such that readers only reviewed each patient’s CT images only once in a given reading session. Total reading time for the study spanned 88 days (07-21-2017 – 10-17-2017) with a median of 21 days [interquartile range: 21 to 25 days] separating reading sessions. Trainees evaluated each case visually on a specially designed computer workstation, circling all liver lesions, and assigning a confidence score (from 0 to 100) for their confidence that the circled lesion represented a hepatic metastases(9).

Pseudo-Reader Study Design Assumptions and Specification

The pseudo-reader approach (Figure 2) is built on the principle that readers are exchangeable. At least two conditions need to be tenable to achieve this. Readers should be similarly trained and demonstrate similar behavior during observer performance studies. For the second consideration, which could be influenced by experience level, consistent use of the confidence rating is one of the most critical aspects. As a standard practice in our MRMC studies, all readers receive detailed instruction on the use of the confidence scale via a standardized set of written and oral instructions, with an intention to standardize readers' use of the confidence scale to minimize the reader-to-reader variation that may occur with inconsistent use of the confidence scale. Both of these assumptions are strong assumptions we considered necessary to validate. To do this, we used data derived from a fully crossed MRMC study to create pseudo-readers and simulated pseudo-reader-based MRMC studies using the following three different strategies.

1. Single Pseudo-Reader. A single pseudo-reader can be derived from the data from a fully crossed MRMC study by randomly sampling one of the reader interpretations from each imaging strategy by patient combination. In the context of this study, there were 10 radiologist interpretations from which to randomly select one representative interpretation for each of the 80 (4 NRM x 20 patient) combinations. Thus, instead of having a 10 reader study with a total of 800 reading interpretations, this pseudo-reader utilized only 80 reading interpretations (one reading for each NRM – patient combination). Note, there are 10^80 possible combinations of single pseudo-readers that can be determined from this study.

2. Performance Stratified Pseudo-Reader. This approach assumes that reader performance varies across readers and that accounting for this variation in the generation of the pseudo-reader will improve precision in the estimation process. To test this concept, a single NRM was selected as if it had been pilot study designed to estimate reader performance figures of merit (FOMs). There are two commonly used approaches for estimating the FOM with a free response paradigm, both generally denoted as jackknife alternative free-response receiver operating characteristic (JAFROC) analysis. The distinction between the two approaches is in how the FOM definitions consider non-localizations (i.e., false positive markings) in positive cases differently. JAFROC1 penalizes the non-localizations in cases with a target lesion whereas JAFROC does not(10). Both FOMs measure subtle differences in reader performance that we desired to account for in our performance estimation process. Accordingly, we constructed a summary composite measure for reader performance that was the mean JAFROC1 and JAFROC reader-specific FOMs. To stratify readers based on performance, the strata were created with the following upper and lower bounds to the mean composite score: [0,.6), [.6, .7), [.7, .8), [.8, .9), and [.9, 1.0). This performance binning resulted in 0, 5, 4, 1, 0, and 0 readers being binned into those categories based on the reader's composite performance on NRM D, respectively. The end result was three pseudo-readers, one with sampling from 5

readers (i.e., those with FOMs in the range [.6, .7)), one with sampling from four readers, and 1 equaling the data from a single reader.

3. Performance Matched Pseudo-Reader. Using the same composite score (i.e., mean of the JAFROC1 and JAFROC FOMs for NRM D) as the performance stratified pseudo-reader, readers were matched in terms of performance based on the rank order of the estimated performance. For example, the mean FOMs for Readers 8 and 6 were 0.61 and 0.62, which were the two lowest FOMs in this study, were paired together. Similarly, readers 9 and 4, 1 and 7, 3 and 10, and 2 and 5 were also paired, resulting in 5 pseudo-readers for the study.

For each of the three potential pseudo-reader specifications, a generalized bootstrap approach was utilized to randomly generate possible study results originating from the observed fully crossed study design. The random selection process was such that if a reader's interpretation for a NRM and patient dataset was selected for incorporation into the pseudo-reader design, all reader marks (lesion localizations and non-lesion localizations), if any, for that case were included as a set. For each potential pseudo-reader design, 2000 replicates were created by randomly selecting a single interpretation for every stratum in the study design. This random selection process within stratum is the manifestation of the exchangeability assumption in the pseudo-reader design. These simulated study results were generated and archived for subsequent analyses.

Mock Study Design

The three pseudo-reader approaches described above were conducive for the general, post-hoc examination of the operating characteristics for each pseudo-reader design; however, the resampling approach used to generate the distribution of FOMs did not directly mimic the use of any of one design for one particular realization of a study's data in practice. If a design were to be implemented prospectively, the results would need to be based only on a single set of reader interpretations. To simulate the results of an actual study, a mock study was created.

Of the three pseudo-reader designs presented above, the performance matched pseudo-reader approach was selected for use in the mock study. This study design resulted in the largest number of pseudo-readers, five, of the designs considered. This would allow for a more precise estimate of the reader variance component in the JAFROC analysis.

As before, NRM D was arbitrarily selected as the pilot study used to estimate reader performance for matching. The composite index defined as the mean JAFROC1 and JAFROC FOMs was used to match readers into pairs. The reader performance on the remaining three NRMs was estimated in two ways. First, the fully crossed results on NRMs A - C that were obtained using all of the original reader marks obtained in the overall study. This fully crossed result was considered the reference. Then, a single performance matched pseudo-reader study was generated using the paired reader data in order to estimate the FOMs for NRMs A – C (i.e., by randomly selecting one of the two paired observations for each reader stratum - NRM – patient combination). The archived bootstrap replicates generated above were used as the sampling frame for the mock study. In particular, one of the 2000 randomly generated pseudo-reader results was to be selected at random to provide

the data for the mock study. For transparency, the following commands were used in R to randomly select the bootstrap replicate that would serve as the study result for the mock study: set.seed(20180816) and head(sample(1:2000), n=1). The randomized selection process resulted in bootstrap replicate number 139 being selected. Note, the seed was set to the numeric date the simulation was run as this is the standard process when randomized reading sets are produced for general MRMC study designs by our team. Accordingly, it was considered the approach that would have been utilized if the mock study had been implemented prospectively. This selection process was also blind to the numeric results of any of the bootstrap replicates.

Statistical Analysis

To summarize human observer performance in the fully crossed study along with all bootstrap replicates generated through the simulation studies, we utilized JAFROC1 FOMs(10). While the utilization of non-localizations in positive cases has been debated(10), this approach was selected since all false positives were of interest and 14 (70%) of the cases had at least one hepatic metastasis. In the context of the bootstrap replicates, the 2000 individual study results were utilized to generate an empirical distribution of FOM estimates. The 95% bootstrap confidence interval was obtained by selecting the 2.5^{th} and 97.5^{th} percentile from the posterior distribution.

These analyses were supported by additional examination of the performance metrics from the mock study. JAFROC1 FOMs were estimated for NRMs A – C for a derived fully crossed design that removed reader marks for NRM D and for the randomly-selected performance matched pseudo-reader study (simulation #139). In addition, lesion detection sensitivity was directly assessed for both study designs. Generalized estimating equations (GEE) estimates of per-lesion sensitivity were estimated. For these estimates, a minimal reader confidence threshold was considered helpful to establish a comparison to clinical practice. For this purpose a reader confidence of >10 was felt to be reasonable (e.g., corresponding to the phrase “probable tiny cysts” that a radiologist might use in a clinical report). Correct localization and task confidence >10 was required for a lesion to be considered detected in the sensitivity calculation.

To understand the reproducibility of confidence scores among readers for all reader interpretations, the intra-class correlation was calculated by constructing datasets that listed each reader-specific confidence scores for each true lesion. If a radiologist detected the lesion, the assigned confidence score was used. If a reader did not detect the lesion, a confidence score of 0 (i.e., the value utilized in a JAFROC1 analysis) was used. In addition, we examined the speed at which cases were reviewed over the course of the study using the internal time stamps recorded by the workstation. Data were grouped into reader-patient combinations and the case reading times were modeled using a random effects model with a fixed effect for reading session number. Post hoc comparisons of the model-based mean times by reading session were estimated. P-values reported are two-sided and have not been adjusted for multiple comparisons. Statistical analysis was conducted using R version 3.4.2 (Vienna, Austria). JAFROC1 FOMs were calculated for every NRM and every reader using the Hillis improvement(11) to the Dorfman, Berbaum and Metz method(12) under the

modeling assumption of random readers - random cases using the RJafroc package v1.0.1. Lesion sensitivity was calculated using the mrmctools package.

Results

Fully Crossed Validation Study Results

The 10 radiologist trainees each read 80 study datasets (4 NRMs on 20 unique patient datasets). Figure 3 plots the estimated FOMs for the JAFROC1 analysis. None of the pairwise comparisons in FOMs among NRMs were statistically significant ($p>0.32$ for the six possible comparisons).The confidence intervals were noticeably and expectedly wide with the limited number of cases examined, and there was significant reader variation in the study. In particular, reader-specific FOMs ranged from 0.601 to 0.796 in the study. The time to read cases using the specialized computer workstation also improved over the study (Figure 4). Like the FOMs, there was significant variability among the readers. The most significant drop in time occurred between the first and second reading sessions.

To assess reproducibility of the confidence scores assigned by the 10 readers to the 33 hepatic metastases, the intra-class correlation coefficient (ICC) was utilized. For NRMs A – D, the ICC (95% CIs) were 0.622 (0.499 – 0.751), 0.604 (0.478 – 0.737), 0.657 (0.538 – 0.778) and 0.609 (0.484 – 0.741), respectively. Based on the common Landis and Koch interpretations(13), this would imply that the inter-reader utilization of the confidence scores was in the substantial agreement range with confidence intervals indicating the potential for moderate agreement.

Pseudo-reader Results

Figure 5 plots the histograms and bootstrap estimates of the FOMs of merits from the three general pseudo-reader strategies. There are two distinct trends in the figure. First, judicious blocking on reader performance provides dramatic improvements (reductions) in the variability of FOMs obtained from a pseudo-reader design. The first row in Figure 5 shows the least reproducibility in findings even though these estimates do in fact align well with the estimate obtained using the fully cross design. The second row demonstrates the performance stratified results. While this provides a more precise estimate of the FOM, there appears to be bias in the estimated result (FOMs were overestimated using the stratification plan studied). This finding is likely a direct result of having too much heterogeneity in the [0.6, 0.7) stratum and heavily weighting a single, strong-performing reader in the [0.7, 0.8) stratum. The effect of performance matching is shown in the bottom row of Figure 5. Here, readers were paired into 5 pseudo-readers. The estimated FOMs were nearly perfectly aligned with the results obtained using the fully crossed study design. It should be noted that the bootstrap confidence interval shown in red in this figure is fundamentally different than the confidence interval presented for the overall JAFROC1 method in this case. Here, the bootstrap interval is a measure of convergence of the pseudo-reader result to the fully crossed study result; not to the general population of all possible FOM results that could be obtained should the fully crossed be replicated.

Mock Study Result

To directly compare the pseudo-reader and fully crossed results, the mock study was created and analyzed as it if were conducted prospectively. Here, unlike the pseudo-reader studies described above, the JAFROC1 estimates of the FOM and associated 95% CIs under both designs were directly comparable. Figure 6 presents the results. The point estimates between the two study designs approaches agree within +/− 3% and the confidence intervals are essentially the same width despite the pseudo-reader approach utilizing half the number of reading interpretations. Table 1 presents the estimate of lesion detection for the fully crossed and performance matched study designs. Similar to the FOMs, there is strong agreement in the estimated lesion sensitivity with numerical matching of the point estimates for two of the three NRMs and NRM B being estimated within +/−3%. The confidence intervals for the pseudo-reader approach, however, were wider, which was a direct result of only utilizing 5 instead of 10 readers to estimate the pooled sensitivity.

Discussion

This study quantified the operating characteristics of human observer performance studies under a new approach to assigning reading sets to human observers by simulating three different pseudo-reader approaches from a fully crossed MRMC dataset. The approach is referred to as a “pseudo-reader”, reflecting the fact that virtual readers are created by combining judiciously assigned datasets to a group of similarly trained readers. This parallelization of the reading list was found to provide comparable results to the traditional fully crossed design. However, unlike the traditional design, the parallelization of the reading across multiple readers has the potential to rapidly evaluate observer performance while directly addressing one of the key limitations in the standard MRMC approach – lesion conspicuity and reader recall. The simulation studies supported the concept that the more exchangeable the readers were, the more the results would align with the fully crossed study.

The logistical and scientific limitation of a large number of reading interpretations has been discussed before with the utilization of a split-plot adaptation of an MRMC study design(4, 5). The work by Obuchowski et al formally developed a test statistic for the split-plot design(4). There are similarities and differences to our proposed approach. The common goal of reducing reader interpretations and accelerating testing is common to the two approaches. There are some fundamental differences in the conceptual approaches to the two study designs. The split-plot design has a formal statistical foundation that builds upon strict nesting of reader pairs within each stratum. In enforcing this hierarchy to the study design reader performance is estimated and affects the variance components. Such a study design is not readily implemented into standard software such as RJafroc and the need for a hierarchical structure to the data limits flexibility with implementation. With the pseudo-reader approach, the reader variation is allowed to coalesce with the residual error. While statistically speaking this could result in a loss of efficiency, it also opens up more flexibility with study designs. As was shown in this study, performance matching on reader’s performance appears important in the conduct of a pseudo-reader design. The pseudo-reader concept thus deviates from the split plot design by measuring and matching readers around

the assumptions of exchangeability of readers. The design is flexible in that many readers may be utilized for a single pseudo-reader in a way that may result in the inability to estimate the variance components attributed to each individual reader. The scenario where there are many new imagining strategies or decision support tools that to be evaluated is expected to be a situation where a pseudo-reader design would be useful. The design could allow for rapid preliminary examination of a wide range of configurations in order to provide some empirical data to plan a confirmatory, standard MRMC study design.

Limitations of the research are worth noting. First, the results of simulations show inherent variability that one might encounter should studies be repeated numerous times. This is a general issue for all MRMC study designs. Confidence intervals help convey this variability. In the case of a pseudo-reader, the within stratum variation attributable to variations in reader performance of matched readers is not directly accounted for in the standard JAFROC analysis. This might suggest that the confidence intervals using pseudo-readers are too narrow (optimistic), but this may be a function of how well the readers are performance matched. Future study is warranted on this topic. Another limitation is that our matched performance approaches are relatively rudimentary, although they showed promise. One could imagine conducting a much more stringent evaluation of reader performance where detection across various lesion sizes was evaluated, confidence score utilization was examined closely, and overall detection ability was quantified through a comprehensive training and evaluation protocol. Our data did not provide this richness, thus computer adaptive approaches to quantify expected reader performance are a topic for further research. In the context of the mock trial, it would have been desirable to have evaluated performance on cases external to the challenge to have provided more purity with respect to the detection task (i.e., it may be possible that recall of conspicuous lesions could confound the results). There is also an inherent limitation in the selection of readers for the study. The readers chosen for the study were relatively homogenous with respect to training and professional experience, something quite different from nay MRMC studies. While there was some consistency in training, reading performance based on the FOMs did demonstrate a range of performance potentially indicating the complexity of the low dose detection tast. These limitations notwithstanding, this was one of the first attempts at validating the pseudo-reader study design.

In conclusion, with attention to reader performance and matching, a multi-pseudo reader, multi-case study design can yield tremendous savings in time while providing comparable levels of quantification of observer performance. Such an approach has the potential to greatly accelerate human evaluation of altered imaging strategies, which is extremely timely given the rapid development of artificial intelligence-based computer decision support tools. In order to achieve this performance gain, the strong assumption of exchangeability of readers needs to be made.

Acknowledgments

Funding: This work was supported by the National Institutes of Health under supplemental award number U01 EB017185. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health.

Abbreviations

AAPM	American Association of Physicists in Medicine
CI	confidence interval
GEE	generalized estimating equations
ICC	intra-class correlation
JAFROC	jackknife alternative free-response receiver operating characteristic
FOM	figures of merit
MRMC	multi-reader-multi case
NRM	noise reduction method

References

1. Thompson JD, Manning DJ, Hogg P. The value of observer performance studies in dose optimization: a focus on free-response receiver operating characteristic methods. J Nucl Med Technol 2013; 41(2):57–64. [PubMed: 23625536]
2. Chakraborty DP. Recent developments in imaging system assessment methodology, FROC analysis and the search model. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 2011; 648:S297–S301.
3. McCollough CH, Bartley AC, Carter RE, et al. Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge. Med Phys 2017; 44(10):e339–e52. [PubMed: 29027235]
4. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. Acad Radiol 2012; 19(12):1508–17. [PubMed: 23122570]
5. Chen W, Gong Q, Gallas BD. Paired split-plot designs of multireader multicase studies. J Med Imaging (Bellingham) 2018; 5(3):031410. [PubMed: 29795776]
6. Low Dose CT Grand Challenge Data Use Agreement Available at: https://www.aapm.org/GrandChallenge/LowDoseCT/documents/DataSharingAgreementLowDoseCTGrandChallenge.pdf. Accessed 09/12/2018.
7. Yu L, Shiung M, Jondal D, McCollough CH. Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols. J Comput Assist Tomogr 2012; 36(4):477–87. [PubMed: 22805680]
8. Fletcher JG, Chen MH, Herman BA, et al. Can radiologist training and testing ensure high performance in CT colonography? Lessons From the National CT Colonography Trial. AJR Am J Roentgenol 2010; 195(1):117–25. [PubMed: 20566804]
9. Fletcher JG, Fidler JL, Venkatesh S, et al. Observer Performance with Varying Radiation Dose and Reconstruction Methods for Detection of Hepatic Metastases. Radiology 2018:180125.
10. Sahiner B, Chakraborty DP, Manning DJ, Yoon H-J. JAFROC analysis revisited: figure-of-merit considerations for human observer studies. Proc SPIE 7263, Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment 2009; 7263:72630T.
11. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in medicine 2007; 26(3):596–619. [PubMed: 16538699]
12. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27(9):723–31. [PubMed: 1399456]
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33(1):159–74. [PubMed: 843571]

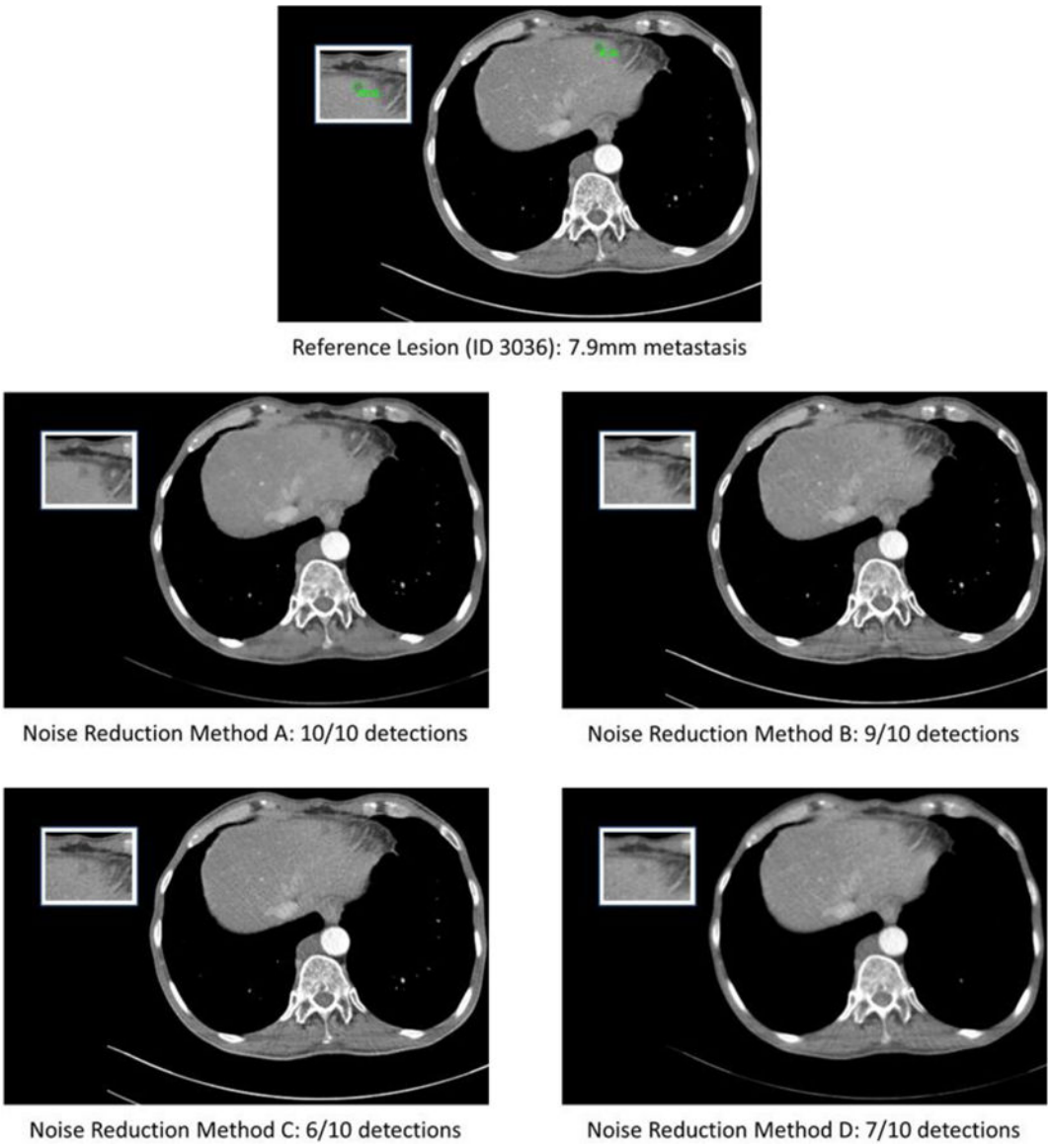

Figure 1.

Representative slices from the four noise reduction methods along with the reference slice. The number of detections out of the 10 readers is noted for each noise reduction method.

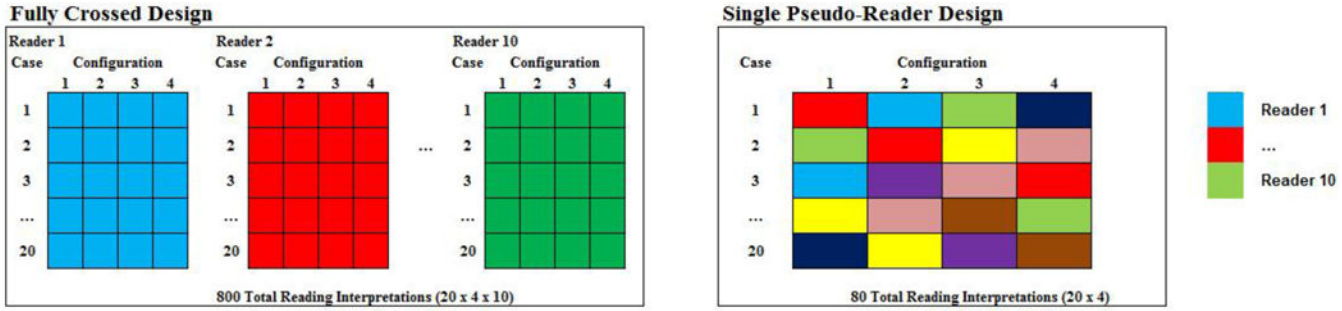

Figure 2.

Representation of how a fully crossed design can be translated into a pseudo-reader study design. Colors in the cells represent individual interpretation by up to 10 readers, as shown. In the fully crossed design, all 10 readers would read the entire panel of 20 cases across the four imaging configurations. In contrast, a pseudo-reader design federates the complete reading coverage over multiple readers resulting in fewer total reading interpretations.

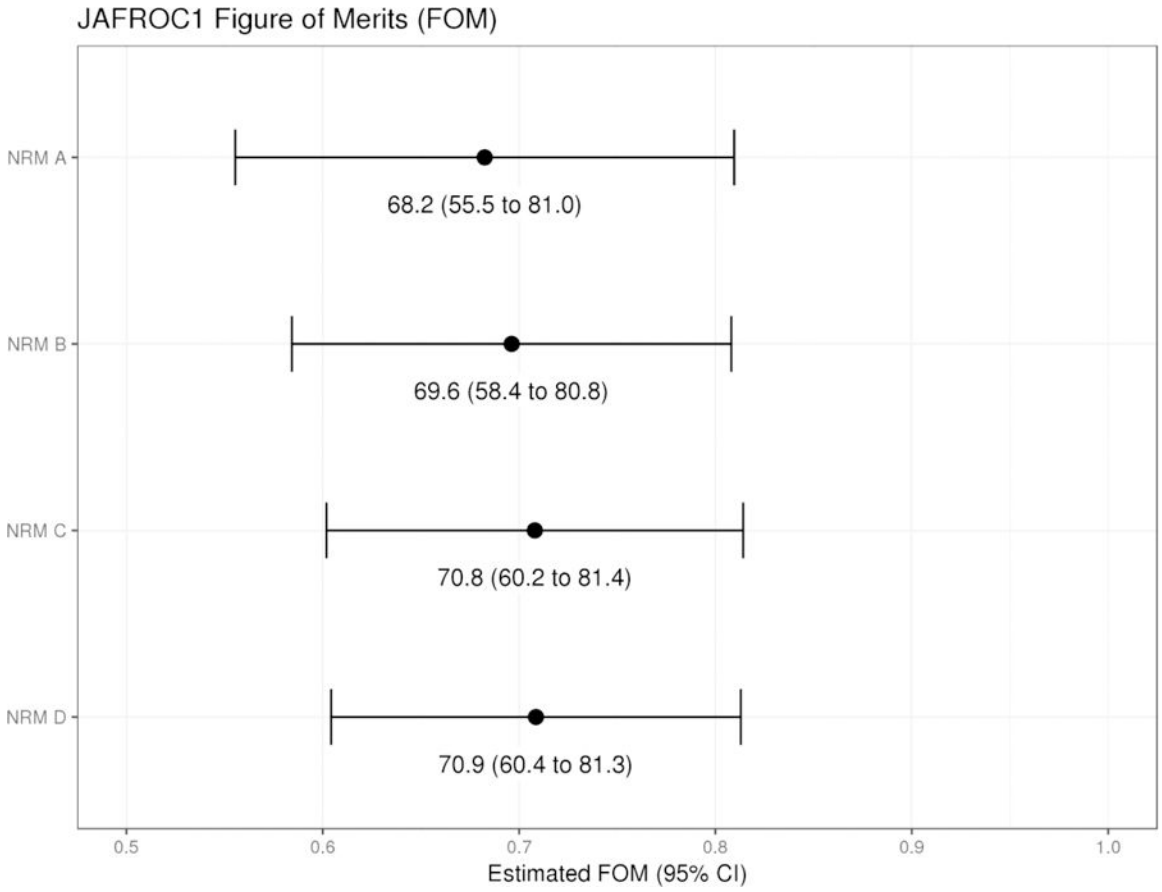

Figure 3.

Results of the fully crossed validation study. A total of 20 patient datasets reconstructed with 4 different noise reduction methods (NRMs) were read by 10 radiologist trainees. The JAFROC figure of merit was estimated using a random-reader, random-case analysis approach that penalized performance for non-localizations ("false positive") in cases with and without true lesions (JAFROC1 analysis).

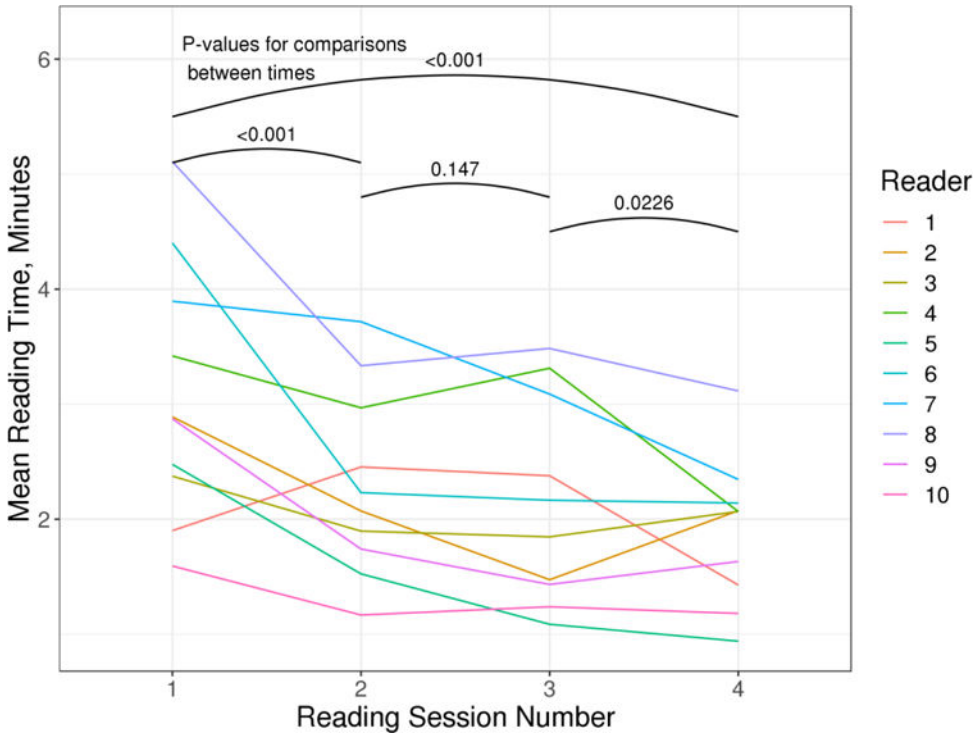

Figure 4.

Longitudinal analysis of the mean case reading times over the four reading sessions stratified by readers. P-values are tests of model-based means between each time point.

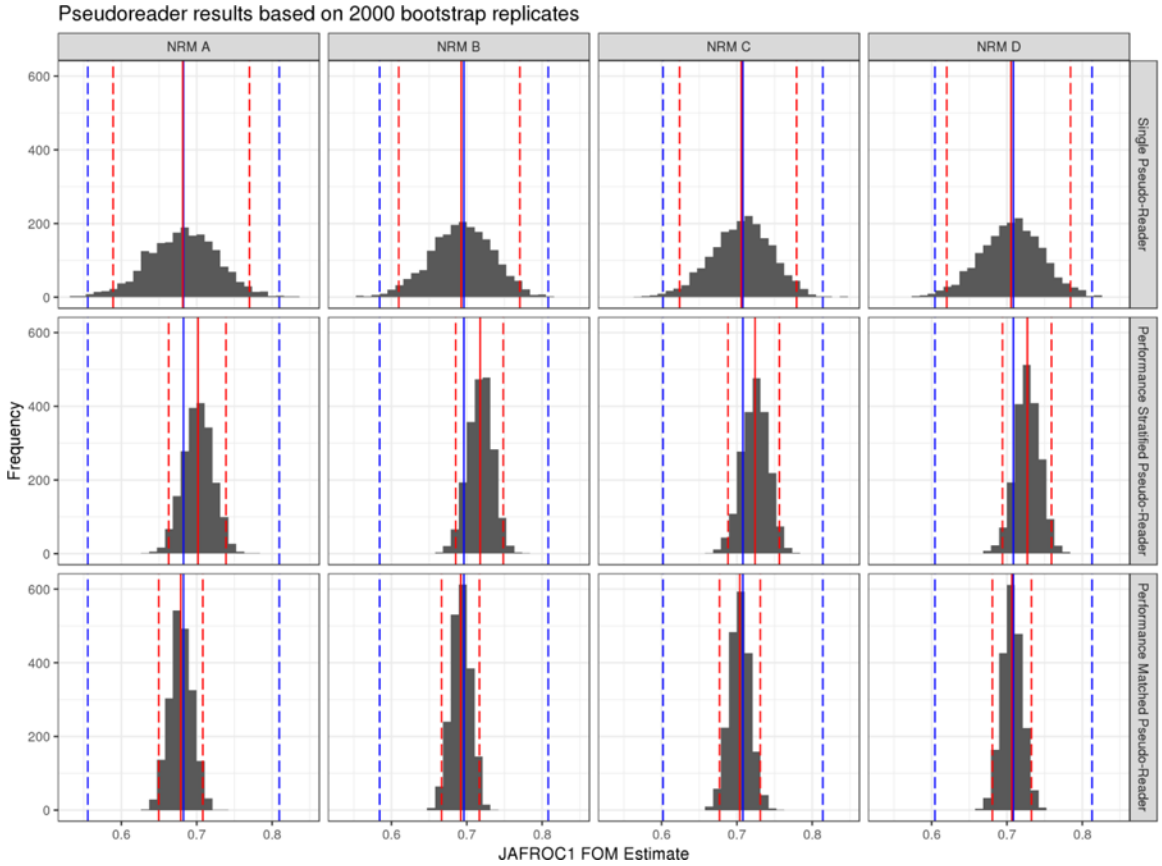

Figure 5.

Histograms of the individual estimates obtained by each of the bootstrap replicates for each pseudo-reader strategy. The blue line (dashed) lines represent the estimated FOM (95% CIs) from the fully crossed study design. The red (dashed) lines represent the mean (95% bootstrap confidence interval) for the pseudo-reader study design. Top row shows the results from a single pseudo-reader generated from the 10 readers. The remaining rows show the effect of stratification (middle row) and matching (bottom row) on estimated FOMs.

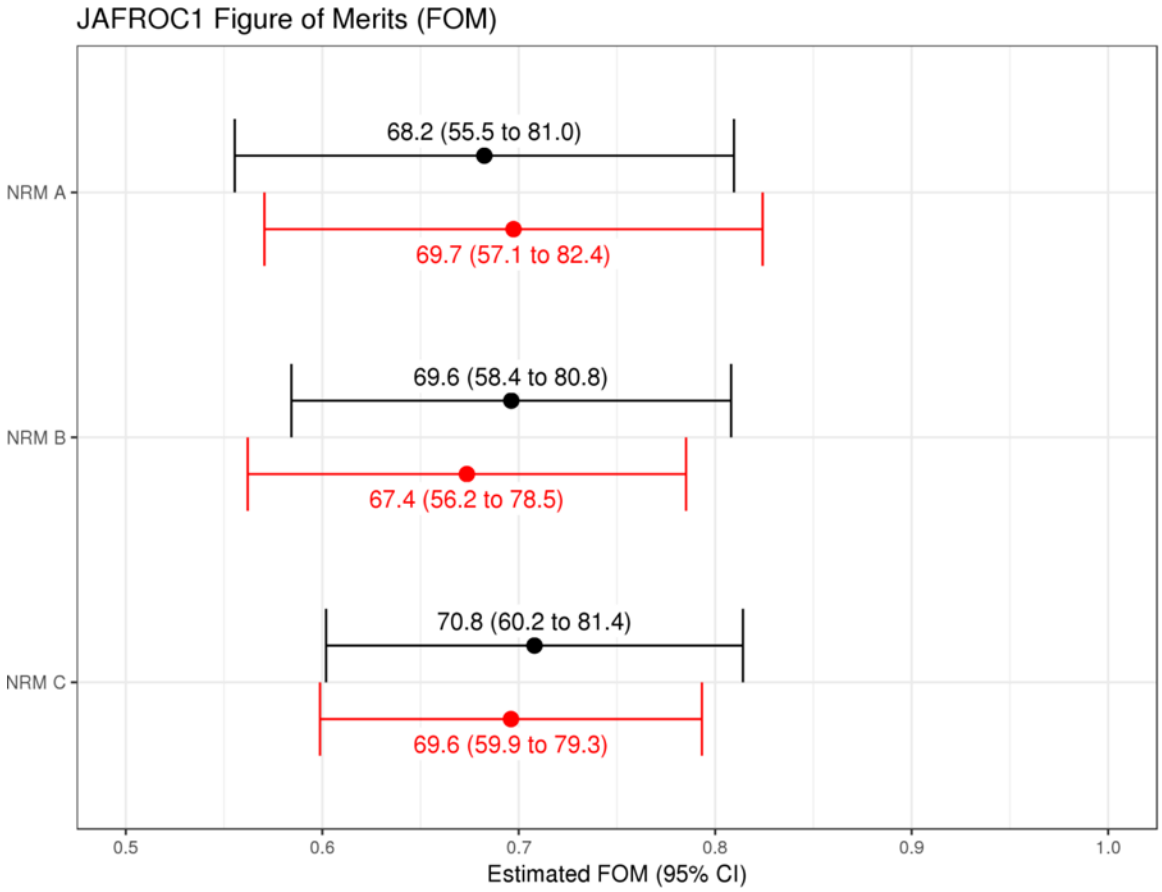

Figure 6.

Comparison of the pseudo-reader result (bottom, red) vs. the fully crossed study (top, black) for a study that utilized noise reduction method D to pair readers based on performance for the mock study. The results of the estimates that utilized 5 pseudo-readers (400 reading interpretations) are compared to those obtained using the fully cross results utilizing all 10 years (800 reading interpretations). For the pseudo-reader estimate, a single estimate from the pseudo-reader was drawn at random from the pool of 2000 bootstrapped estimates.

Table 1.

Lesion-specific sensitivity for the detection of 33 lesions among 20 patient datasets. The estimated sensitivity and confidence intervals are based on GEEs that account for the repeated interpretations of the patient datasets by either 10 readers or 5 performance-matched pseudo-readers derived from the fully crossed study design. The range presented is for the individual reader or pseudo-reader performances.

	Fully Crossed Design (10 readers)		Performance Matched Pseudo-Readers (5 pseudo-readers)	
Noise Reduction Method	**Sensitivity (95% GEE CI)**	**Range**	**Sensitivity (95% GEE CI)**	**Range**
A	72.1 (67.3 to 77.0)	57.6 to 81.8	72.1 (65.3 to 79.0)	66.7 to 78.8
B	61.5 (56.3 to 66.8)	45.5 to 75.8	60.0 (52.5 to 67.5)	51.5 to 69.7
C	58.2 (52.9 to 63.5)	45.5 to 72.7	58.2 (50.7 to 65.7)	48.5 to 72.7