Published in final edited form as:

Acad Radiol. 2008 December ; 15(12): 1567–1573. doi:10.1016/j.acra.2008.07.011.

Agreement of the order of overall performance levels under different reading paradigms

David Gur, ScD¹, Andriy I. Bandos, PhD², Amy H. Klym, MS¹, Cathy S. Cohen, MD¹, Christiane M. Hakim, MD¹, Lara A. Hardesty, MD³, Marie A. Ganott, MD¹, Ronald L. Perrin, MD¹, William R. Poller, MD⁴, Ratan Shah, MD¹, Jules H. Sumkin, DO¹, Luisa P. Wallace, MD¹, and Howard E. Rockette, PhD²

¹ University of Pittsburgh, Department of Radiology

² University of Pittsburgh, Department of Biostatistics

³ University of Colorado Hospital Breast Center

⁴ West Penn Allegheny Health System

Abstract

Rationale and Objectives—To investigate consistency of the orders of performance levels when interpreting mammograms under three different reading paradigms.

Materials and Methods—We performed a retrospective observer study in which nine experienced radiologists rated an enriched set of mammography examinations that they personally had read in the clinic (“individualized”) mixed with a set that none of them had read in the clinic (“common set”). Examinations were interpreted under three different reading paradigms: binary using screening Breast Imaging Reporting and Data System (BI-RADS), receiver operating characteristic (ROC), and free-response ROC (FROC). The performance in discriminating between cancer and non-cancer findings under each of the paradigms were summarized using Youden’s index/2+0.5 (Binary), nonparametric area under the ROC curve (AUC), and an overall FROC index (JAFROC-2). Pearson correlation coefficients were then computed to assess consistency in the ordering of observers’ performance levels. Statistical significance of the computed correlation coefficients was assessed using bootstrap confidence intervals obtained by re-sampling sets of examination-specific observations.

Results—All but one of the computed pair-wise correlation coefficients were larger than 0.66 and were significantly different from zero. The correlation between the overall performance measures under the Binary and ROC paradigms was the lowest (0.43) and was not significantly different from zero (95% CI: from −0.078 to 0.733).

Conclusion—The use of different evaluation paradigms in the laboratory tends to lead to consistent ordering of the overall performance levels of observers. However, one should recognize that conceptually similar performance indices resulting from different paradigms often measure different performance characteristics and thus disagreements are not only possible but frequently quite natural.

Correspondence: David Gur, University of Pittsburgh, Department of Radiology, Imaging Research, 3362 Fifth Avenue, Pittsburgh, PA 15213-3180, Phone :412-641-2513, Fax: 412-641-2582, Email: gurd@upmc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

observer performance; receiver operating characteristic (ROC); free-response ROC (FROC); binary ratings

Introduction

Considerable progress has been made in the last two decades with regard to our understanding of the methodologies needed for the analysis of retrospective observer performance studies [1–10]. Perhaps the simplest type of performance study is a binary detection task (binary response) which primarily provides performance measures in terms of sensitivity and specificity. Another frequently used approach is a Receiver Operating Characteristic (ROC) paradigm that provides information on how sensitivity and specificity change with varying decision thresholds for multi-category (discrete or semi-continuous) type ratings [11–14]. Another approach that is recently gaining acceptance is the Free-Response ROC (FROC) paradigm that uses a “locate then rate” approach with an a priori unknown number of suspected locations per image or examination [15]. The primary purpose of a specific study should determine which of the approaches is optimal in terms of validity, appropriateness, and clinical relevance [16–18]. Unfortunately, specific paradigms are frequently selected based on prior experience and/or expertise of the investigators, available resources, and prior studies performed to address the same or similar hypotheses.

Since the three paradigms can often be implemented in the same situation, an important question of interest is whether or not the results of a study performed under one paradigm (e.g. ROC) are likely to remain had the study been performed under another paradigm (e.g. FROC). Different observers may behave substantially differently under different paradigms and there are virtually no data to support the consistency of radiologists’ performance levels when interpreting examinations under these different paradigms. It is a well known and studied phenomenon that radiologists’ interpretations vary substantially leading to significant intra- and inter observer variability in general [19,20] and during screening mammography interpretations in particular [21–26]. Thus, it is important to investigate whether the widely varying performance levels of readers under different paradigms correlate with each other.

An underlying assumption for all approaches to design and analysis of observer performance studies is that each reader operates in some possibly complex but self-consistent manner in determining whether the image in question actually depicts (or not) a well defined abnormality and the possible depiction exceeds a predetermined level of suspicion (“threshold”) warranting a recommendation for further action [15,27]. There are studies that suggest there should be little difference between using discrete or semi-continuous rating scales in ROC studies [28, 29]. However, in principle, there are no fundamental underlying reasons why different paradigms, even if they differ only with respect to the rating scale, could not produce different results [17].

In our previous paper [30] we experimentally demonstrated that reader-specific performances under the Binary paradigm do not exhibit systematic fluctuations from the ROC curves. That result experimentally supports the theoretical relationship often presumed to exist, at least on average, between the ROC curve and the operating point (sensitivity, specificity) estimated under the ROC and Binary paradigms. However, the evaluation of system’s performance levels is often conducted with overall summary indices such as Youden’s index/2+0.5 (Binary) [31], area under the ROC curve (AUC) [32] or jackknife FROC (JAFROC-2) [33]. Although these indices are conceptually similar, namely all three of them describe the ability of the system to discriminate between negative and positive findings; they depict scale-specific

performance measures and quantitatively can be quite different even for the same diagnostic system. Hence, a direct comparison of the overall performance indices between the paradigms could potentially be meaningless. Yet, because different paradigms are used to investigate the same or similar problems, it is important to know whether comparisons made under one paradigm tend to agree with comparisons made under another paradigm. Therefore, as a part of a comprehensive retrospective study designed to investigate several aspects related to the “laboratory effect”, if any, we investigated the consistency in the order of reader-specific overall performance levels, if any, during the interpretation of the same examinations when read under the Binary, ROC, and FROC paradigms.

Materials and Methods

General Study Design

We focus our analysis on data collected during an observer performance study in which 9 board-certified Mammography Quality Standards Act (MQSA) qualified radiologists interpreted a set of 155 screen-film mammography (SFM) examinations they had not read in the clinic enriched with examinations that they had individually interpreted in the clinic between 2 and 6 years previously. The methodology employed during this observer performance study, which was primarily designed to evaluate a “laboratory effect”, if any, has been described in detail elsewhere [34]. In brief, each radiologist independently interpreted examinations as they would in the clinic, rating the right and left breast separately in three reading modes over a period of 20 months under a mode balanced design. The three modes included a “clinical – BI-RADS “ screening rating mode using the screening Breast Imaging Reporting and Data System (BI-RADS) recommendations [35], an ROC rating mode with abnormality presence probability rating scale of 0–100, and a FROC mode [15]. All nine radiologists read the “common” set of 155 SFM examinations originally read in the clinic by other radiologists not participating in the study and an “individualized” set of examinations that had been clinically read by him/her between 2 and 6 years previously. “Common” and “individualized” examinations were mixed and radiologists read all examinations in one mode before moving to the next one. The study included a pre-determined enriched distribution of pathology confirmed cancers (i.e., actually positive examinations depicting either masses, micro-calcifications clusters, or both) and actually negative examinations. Among the total of 1,367 examinations selected for this study 354 (25.9%) examinations depicted verified cancers. In total there were 875 (354+521) examinations depicting at least one abnormality (verified cancer or benign), 522 (59.7%) depicted masses only, 276 (31.5%) depicted micro-calcification clusters, and the remaining 77 (8.8%) depicted both masses and micro-calcification clusters. The remaining 492 examinations were negative. If there was a comparison examination (2 or more years prior when available, or one year prior when the only available examination) used during the original clinical interpretation it was made available during the interpretations in the laboratory.

During the clinical – BI-RADS mode observers were first presented with a choice of rating the examination of each breast as “negative”, “benign”, or recommended for “recall”. If a “benign” or a “recall” rating was entered, observers were asked to identify the type of abnormality(s) in question (i.e. “mass”, “micro-calcifications”, “other”) and in case of a recommendation for “recall” to select one or more recommended procedures (e.g., spot CC/spot 90, spot CC/whole breast). During the ROC mode radiologists separately rated their assessment of the likelihood that either masses or micro-calcification clusters (or both) were depicted on each breast using a 101 category rating scale (0% – 100%). For any rating greater than 4% a second slider appeared specific for that abnormality and “assuming that the abnormality in question was actually present” readers rated their assessment of the likelihood (0%–100%) that the suspected abnormality in question depicts an actual cancer. In the FROC mode, readers were presented

with a schematic drawing of the four film views (i.e., right cranio-caudal (CC), right mediolateral-oblique (MLO), left CC, and left MLO), and the reader could click on any location of any view to “locate” a suspicious region followed by a query to select the abnormality in question (i.e. “mass”, “micro-calcifications”, or “other”) and a slider to rate the likelihood (0%–100%) that the suspected abnormality in question is actually a cancer. This was followed by a question if the abnormality was depicted on the other ipso-lateral view (“yes/no”). If the answer to the latter was “yes” the reader had to mark the ipsilateral image and provide a rating. Multiple locations and abnormality types could be marked on each image (including none), view and/or examination as the reader deemed appropriate (i.e. free response). If no abnormality was suspected, (i.e. “definitely negative”) the reader could just click “done”. The location “truth” was marked by an experienced interpreter guided by the diagnostic and biopsy reports of location. A second experienced reader was consulted in all cases of less certainty about the location of an abnormality depicted on either one or both views. Consensus agreement was achieved in all cases selected for the study.

Readers were not restricted with respect to time spent interpreting each examination or the number of examinations read during a session. One mode was completed prior to beginning the next one after a predetermined time delay. The study was mode balanced, and all readers continued their routine clinical practice in women imaging during the study. All markings and ratings were recorded electronically in a database.

Data Analyses

A breast-based rating analysis was performed. A breast depicting a verified cancer is treated as an actually positive unit, and a breast without confirmed cancer during the examination in question and for one year thereafter is treated as an actually negative unit.

For the clinical – BI-RADS mode we computed sensitivity and specificity for each radiologist. We define sensitivity here as the fraction of the recalled breasts out of all breasts depicting verified cancers and specificity is defined as the fraction of not recalled breasts out of all verified “cancer free” breasts. The overall performance of each radiologist was summarized using $(\text{sensitivity} + \text{specificity})/2$ or (Youden’s index/2+0.5) [36], which is equivalent to the area under the ROC “curve” obtained by connecting the operating point (sensitivity, specificity) to the (0,0) point and the (1,1) point with straight line segments [31].

In the ROC mode the sensitivity and specificity for each rating threshold were defined similarly to the Binary mode. The overall performance of each radiologist was estimated using the nonparametric AUC [32].

In the FROC mode a mark is considered to be True Positive if the following conditions are satisfied: 1) the center of the mark is located within the “acceptance” distance (or target) of the predetermined “center of the verified cancer” and 2) the type of abnormality recorded by the radiologist for the specific mark coincides with that of the targeted abnormality. We reduced the collected FROC data to the marks-per-breast level by combining the information from the CC and MLO views. The detection rating for the specific abnormality being suspected was taken as the maximum of the two detection ratings, and it was considered a “True Positive” mark if either or both of the marks on the two views was a “True Positive”. The overall performance for the FROC mode was assessed with the JAFROC-2 index [33] which, similarly to the AUCs under other paradigms, can be interpreted as the probability that the “most suspicious” mark on a verified negative or “cancer-free” breast is less suspicious (i.e. lower rating) than a True Positive mark (indicating a cancer) on a breast with a verified cancer. We also analyzed JAFROC-2 indices obtained from the FROC data preprocessed with varying acceptance radii of 20, 30 and 40 pixels on displayed schematic breast drawings.

To assess whether radiologists who performed better using these overall summary indices under one paradigm also performed better under the other paradigms, we computed the Pearson correlation coefficients between overall performance summaries for each pair of the assessment paradigms. Pearson correlation coefficients tend to up-weight disagreements between substantially differing performances. To assess whether the observed value of the correlation could have been caused by random fluctuations in the data we used 95% bootstrap confidence intervals. In order to account for the correlation due to the use of the same examinations and for correlation between the evaluations of the two breasts of the same woman we bootstrapped (re-sampled with replacement) all observations pertinent to the evaluation of a specific examination. In this procedure, examinations with and without verified cancers, as well as examinations from common and reader-specific groups were re-sampled separately.

Results

Table 1 shows the estimated performance levels under the different paradigms for each reader and the average over all nine readers. For individual readers Youden's index/ $2+0.5$ ranged from 0.709 to 0.812 with an average of 0.770. AUC ranged from 0.729 to 0.865 with an average of 0.820. The average JAFROC-2 over all readers were 0.532 (0.425 to 0.598), 0.633 (0.537 to 0.707) and 0.695 (0.591 to 0.746) for the three acceptance target sizes (i.e. 20, 30 and 40 pixels in acceptance target radii), respectively.

Table 2 provides the computed Pearson correlation coefficients and their 95% bootstrap confidence intervals for each of the considered evaluation paradigms. The correlations between the overall performance measures of the Binary and FROC, and ROC and FROC paradigms ranged from 0.67 to 0.83. Their 95% bootstrap confidence limits are substantially far from zero with the closest point being 0.184 for the correlation of Binary and FROC with the smallest acceptance target. It is worth noting that the 99.5% confidence intervals for the correlations between Binary and FROC, and ROC and FROC paradigms also do not include zero. The correlation between ROC and FROC performances tend to be higher than the correlation between Binary and FROC performances. The correlation between the overall performance indices of the Binary (Youden's/ $2+0.5$) and ROC (AUC) paradigms is relatively low (0.43). The 95% confidence interval includes zero (-0.078, 0.733).

Discussion

Our study shows that during a retrospective laboratory study, the use of the Binary versus FROC, or ROC versus FROC reading paradigms tend to result in consistent relative performance levels by radiologists suggesting that comparisons made under different modes should hold in principle, regardless of the paradigm selected by the investigator. However, despite the conceptual similarity of the overall performance indices these summary indices actually measure quite different performance characteristics. The performance characteristics under the Binary paradigm require correct decisions whether to recall or not to recall the women for additional diagnostic follow ups. Under the ROC paradigm good performance also requires a consistency in the rank ordering of multiple subjects. Under the FROC paradigm there is an additional requirement, namely, a correct detection of all abnormal locations within each subject. The overall performance indices under the Binary (Youden's index/ $2+0.5$) and ROC (AUC) paradigm represent but one example of the possible discrepancies between these indices.

In a prior paper [30] we demonstrated an absence of systematic differences between the radiologist-specific Binary "operating point" and his/her ROC curve. That seemingly contradicting finding to the results of this paper is actually not a contradiction, rather a natural consequence of the differences in the performance measures used under the different

paradigms. The appearance of a disagreement is primarily the result of the “between-reader” variability in measured specificities under the Binary paradigm. Figure 1 demonstrates a hypothetical curve where the performance (solid line and $AUC = 0.82$) is uniformly inferior to a second curve (dashed lines and $AUC = 0.85$) under the ROC paradigm. However, as a result of the differences in specificity levels for the selected operating points on the curve, the ROC curve with the lower AUC has a greater overall performance under the Binary paradigm (Youden’s index/2+0.5 is 0.75 vs. 0.73). The parameters of the example were specifically selected from the range of the summary indices computed from the data analyzed in this paper.

We also observed the theoretically expected increase of readers’ performances under the FROC paradigm when the acceptance target sizes increased [37]. Furthermore, the correlation between the overall performance measures of the Binary or the ROC with the FROC increases with increasing size of the “acceptance target”. The observed phenomenon is in agreement with theoretical considerations. Indeed, in a general case, when the acceptance radius increases to the level where a single mark covers the entire image the FROC data is reduced to ROC-type data, and, assuming that artificially combining the ratings of multiple marks on an image is consistent with the considerations actually made by the observer, we could expect the correlation to be very high. Although the consideration made by observers when rating an examination are likely to be substantially more complex than the above implied approach of artificially combining marks and using the highest rating, the general tendency is likely to be preserved.

We emphasize that the actual study design and rating approach should be determined by the specific question being investigated and the expected relevance of the results to the actual clinical practice. Although the consistency expected from the results of an observer performance study when using any of the three rating approaches presented here suggest that the results of comparisons of systems under one paradigm tend to agree with comparisons under another paradigm, in reality the results could be different. As important to note, the actual validity or generalizeability of results from retrospective laboratory studies to the clinical environment is beyond the scope of this paper.

In general, specificity levels in this study were low (recall rates were high) because our case sets included a large fraction of examinations depicting benign findings and a substantial number of these could have been recalled for verification reasons in the actual clinical environment, as well.

In summary, we can expect that comparisons made between systems as a result of using any of the three paradigms should on average be reasonably consistent. The agreement is especially strong between conceptually similar paradigms such as ROC and FROC with large acceptance targets. However, an investigator should always be cognizant of the fact that overall performance measures under different paradigms measure different characteristics and there may be situations where the readers or systems are substantially different under one paradigm and are practically indistinguishable or even reversely ordered under another paradigm.

Acknowledgements

Funding: This work is supported in part by Grants EB006388, EB003503, and EB002106 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institute of Health.

References

1. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837–845. [PubMed: 3203132]
2. 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):723–731. [PubMed: 1399456]
3. Obuchowski NA, Rockette HE. Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: an ANOVA approach with dependent observations. *Communications Statistics Simulations Computations* 1995;24:285–308.
4. Beiden SV, Wagner RF, Campbell G. Components of variance models and multiple bootstrap experiments: An alternative method for random effects, receiver operating characteristics analysis. *Acad Radiol* 2000;7:341–349. [PubMed: 10803614]
5. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *The Canadian Journal of Statistics* 2000;28:731–750.
6. Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. *Acad Radiol* 2002;9(11):1264–1277. [PubMed: 12449359]
7. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Acad Radiol* 2004;11(9):980–995. [PubMed: 15350579]
8. Bandos AI, Rockette HE, Gur D. A permutation test for comparing ROC curves in multireader studies. *Acad Radiol* 2006;13:414–420. [PubMed: 16554220]
9. Gallas B. One-shot estimate of MRMC variance: AUC. *Acad Radiol* 2006;13:353–362. [PubMed: 16488848]
10. Bandos AI, Rockette HE, Gur D. Exact bootstrap variances of the area under the ROC curve. *Communications in Statistics – Theory & Methods* 2007;36(13):2443–2461.
11. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. *Radiology* 2003;228(1):10–14. [PubMed: 12832568]
12. Shah SK, McNitt-Gray MF, De Zoysa KR, et al. Solitary pulmonary nodule diagnosis on CT: results of an observer study. *Acad Radiol* 2005;12(4):496–501. [PubMed: 15831424]
13. Skaane P, Balleyguier C, Diekmann F, et al. Breast Lesion Detection and Classification: Comparison of Screen-Film Mammography and Full-Field Digital Mammography with Soft-copy Reading--Observer Performance Study. *Radiology* 2005;237(1):37–44. [PubMed: 16100086]
14. Shiraishi J, Abe H, Li F, Engelmann R, MacMahon H, Doi K. Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs ROC analysis of radiologists' performance. *Acad Radiol* 2006;13(8):995–1003. [PubMed: 16843852]
15. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol* 2007;14(6):723–748. [PubMed: 17502262]
16. Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur D. A comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. *Med Phys* 2005;32(4):1031–4. [PubMed: 15895587]
17. Gur D, Rockette HE, Bandos AI. "Binary" and "Non-Binary" Detection Tasks: Are Current Performance Measures Optimal? *Acad Radiol* 2007;14(7):871–876. [PubMed: 17626312]
18. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Acad Radiol* 2001;8(4):328–334. [PubMed: 11293781]
19. Metz CE, Shen JH. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. *Med Decis Making* 1992;12(1):60–75. [PubMed: 1538634]
20. Swensson RG, King JL, Good WF, Gur D. Observer variation and the performance accuracy gained by averaging ratings of abnormality. *Med Phys* 2000;27(8):1920–33. [PubMed: 10984238]
21. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med* 1994;331(22):1493–1499. [PubMed: 7969300]

22. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Arch Intern Med* 1996;156(2):209–213. [PubMed: 8546556]
23. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? *J Women's Health* 1998;7:443–449.
24. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. *J Natl Cancer Inst* 2002;94:369–375. [PubMed: 11880475]
25. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst* 2003;95(4):282–290. [PubMed: 12591984]
26. Beam CA, Conant EF, Sickles EA. Factors affecting radiologists inconsistency in screening mammography. *Acad Radiol* 2002;9:531–540. [PubMed: 12458879]
27. Eng J. Receiver operating characteristic analysis: a primer. *Acad Radiol* 2005;12(7):909–916. [PubMed: 16039544]
28. Berbaum KS, Dorfman DD, Franken EA Jr, Caldwell RT. An empirical comparison of discrete ratings and subjective probability ratings. *Acad Radiol* 2002;9(7):756–763. [PubMed: 12139089]
29. Rockette HE, Gur D, Metz CE. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Investigative Radiology* 1992;27(2):169–172. [PubMed: 1601610]
30. Gur D, Bandos AI, King JL, et al. Binary and Multi-Category Ratings in a Laboratory Observer Performance Study: A Comparison. *Med Phys*. submitted
31. Hilden J. Regret graphs, diagnostic uncertainty and Youden's index. *Statistics in Medicine* 1996;15:969–986. [PubMed: 8783436]
32. Hanley JA, McNeil BJ. The meaning and use of the Area under Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982;143:29–36. [PubMed: 7063747]
33. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys* 2004;31(8):2313–2330. [PubMed: 15377098]
34. Gur D, Bandos AI, Cohen CS, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Clinical Prospective and Laboratory Mammography Interpretations. *Radiology*. in press
35. American College of Radiology (ACR). Breast imaging reporting and data system atlas (BI-RADS atlas). Reston (VA): American College of Radiology; 2003 [Accessed October 1, 2003]. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/BIRADSAtlas.aspx
36. Youden WJ. An index for rating diagnostic tests. *Cancer* 1950;3:32–35. [PubMed: 15405679]
37. Chakraborty D, Yoon HJ, Mello-Thoms C. Spatial localization accuracy of radiologists in free-response studies: Inferring perceptual FROC curves from mark-rating data. *Acad Radiol* 2007;14(1):4–18. [PubMed: 17178361]

Figure 1. An example of discordant ordering of overall performance measures under the Binary and ROC paradigms in the presence of perfect “agreement” of actual performances under the two paradigms

The area under the solid ROC curve (a) is 0.82. The point on this curve (b) has coordinates (0.35, 0.85) corresponding to a Youden’s index+1)/2 of 0.75. The area under the dashed ROC curve (c) is 0.85. The point on this curve (d) has coordinates (0.14, 0.60) which correspond to the (Youden’s index+1)/2 of 0.73.

Table 1
Binary, ROC, and FROC performance measures for each of the readers for the combined set of “common” and “individualized” examinations.

Readers	Number of women without verified cancer		Number of women with verified cancer		Actually negative breasts (without cancer)		Actually positive breasts (with cancer)		AUC-type Summary indices under different paradigms		
	Total (=individualized + common)	Total (=individualized + common)	Total (=individualized + common)	Total (=individualized + common)	AUC ((Se+Sp)/2)	AUC	(R=20)	JA-FROC2(R=30)	Binary (R=40)	ROC	FROC
1	191 (=100 + 91)	106 (=42 + 64)	488 (=242 + 246)	106 (=42 + 64)	0.770	0.859	0.574	0.707	0.739		
2	198 (=107 + 91)	97 (=33 + 64)	492 (=246 + 246)	98 (=34 + 64)	0.709	0.813	0.425	0.537	0.617		
3	187 (=96 + 91)	113 (=49 + 64)	487 (=241 + 246)	113 (=49 + 64)	0.782	0.818	0.540	0.605	0.683		
4	190 (=99 + 91)	104 (=40 + 64)	483 (=237 + 246)	105 (=41 + 64)	0.807	0.805	0.538	0.659	0.723		
5	188 (=97 + 91)	108 (=44 + 64)	484 (=238 + 246)	108 (=44 + 64)	0.768	0.865	0.598	0.675	0.746		
6	194 (=103 + 91)	82 (=18 + 64)	470 (=224 + 246)	82 (=18 + 64)	0.740	0.729	0.444	0.544	0.591		
7	187 (=96 + 91)	88 (=24 + 64)	462 (=216 + 246)	88 (=24 + 64)	0.790	0.842	0.549	0.639	0.724		
8	199 (=108 + 91)	78 (=14 + 64)	476 (=230 + 246)	78 (=14 + 64)	0.748	0.797	0.557	0.649	0.692		
9	207 (=116 + 91)	90 (=26 + 64)	504 (=258 + 246)	90 (=26 + 64)	0.812	0.854	0.560	0.686	0.740		
Average					0.770	0.820	0.532	0.633	0.695		

Table 2

Computed Pearson correlation coefficients for all readers using overall performance measures computed for ratings under the different paradigms.

Scales of pairwise correlated measures		Pearson correlation coefficient	
First scale	Second scale	Estimate	95% Bootstrap confidence interval
<i>Binary</i>	<i>ROC</i>	0.430	(−0.078, 0.733)
	<i>FROC (R=20)</i>	0.667	(0.184, 0.836)
	<i>FROC (R=30)</i>	0.694	(0.235, 0.841)
	<i>FROC (R=40)</i>	0.750	(0.262, 0.890)
<i>ROC</i>	<i>FROC (R=20)</i>	0.711	(0.275, 0.875)
	<i>FROC (R=30)</i>	0.726	(0.345, 0.877)
	<i>FROC (R=40)</i>	0.834	(0.485, 0.928)
	<i>FROC (R=30)</i>	0.924	(0.741, 0.967)
<i>FROC (R=20)</i>	<i>FROC (R=40)</i>	0.935	(0.703, 0.969)
<i>FROC (R=30)</i>	<i>FROC (R=40)</i>	0.951	(0.827, 0.977)