The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations[1]

David Gur, ScD
Andriy I. Bandos, PhD
Cathy S. Cohen, MD
Christiane M. Hakim, MD
Lara A. Hardesty, MD
Marie A. Ganott, MD
Ronald L. Perrin, MD
William R. Poller, MD
Ratan Shah, MD
Jules H. Sumkin, DO
Luisa P. Wallace, MD
Howard E. Rockette, PhD

Purpose: To compare radiologists' performance during interpretation of screening mammograms in the clinic with their performance when reading the same mammograms in a retrospective laboratory study.

Materials and Methods: This study was conducted under an institutional review board–approved, HIPAA-compliant protocol; the need for informed consent was waived. Nine experienced radiologists rated an enriched set of mammograms that they had personally read in the clinic (the "reader-specific" set) mixed with an enriched "common" set of mammograms that none of the participants had previously read in the clinic by using a screening Breast Imaging Reporting and Data System (BI-RADS) rating scale. The original clinical recommendations to recall the women for a diagnostic work-up, for both reader-specific and common sets, were compared with their recommendations during the retrospective experiment. The results are presented in terms of reader-specific and group-averaged sensitivity and specificity levels and the dispersion (spread) of reader-specific performance estimates.

Results: On average, the radiologists' performance was significantly better in the clinic than in the laboratory ($P = .035$). Interreader dispersion of the computed performance levels was significantly lower during the clinical interpretations ($P < .01$).

Conclusion: Retrospective laboratory experiments may not represent either expected performance levels or interreader variability during clinical interpretations of the same set of mammograms in the clinical environment well.

[1] From the Departments of Radiology (D.G., C.S.C., C.M.H., M.A.G., R.L.P., R.S., J.H.S., L.P.W.) and Biostatistics (A.I.B., H.E.R.), University of Pittsburgh School of Medicine, 3362 Fifth Ave, Pittsburgh, Pa 15213-31803; University of Colorado Hospital Breast Center, Aurora, Colo (L.A.H.); and West Penn Allegheny Health System, Pittsburgh, Pa (W.R.P.). Received November 19, 2007; revision requested December 21, 2007; final revision received January 17, 2008; accepted March 28. Supported in part by grants EB001694 and EB003503 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. **Address correspondence to** D.G. (e-mail: *gurd@upmc.edu*).

Important progress has been made in our understanding of the use of retrospective observer performance studies in the evaluation of diagnostic imaging technologies and clinical practices as well as the methods needed for the analysis of such studies (1–8). A frequently used approach is a receiver operating characteristic (ROC)–type study that provides information about how sensitivity varies as specificity changes while accounting for reader and case variability (9–12).

The most relevant question of interest in all of these studies is not whether the results can be generalized to cases, readers, abnormalities, and modalities under the study conditions, but rather whether the results of a given study lead to valid inferences on the potential effect of different technologies or practices in the actual clinical environment. Experimental conditions that are required in the vast majority of observer performance studies could affect human behavior in a manner that would limit the clinical relevance of inferences made (13). Data have been collected in the attempt to assess the possibility of a "laboratory effect" in observer performance studies and how it could affect the generalizeability of results (14).

Because large observer variability has been reported in many studies, in particular during the interpretation of mammograms (15–19), we performed a comprehensive, large observer study designed to compare radiologists' performance during the interpretation of screening mammograms in the clinic to their performance when reading the same images in a retrospective laboratory study.

Materials and Methods

General Study Design

Nine board-certified, Mammography Quality Standards Act–qualified radiologists (with 6–32 years of experience in interpreting breast imaging studies and who perform more than 3000 breast examinations per year) were selected to participate in the study on the basis of the number of screening mammograms read during the period from which we selected the images. Each reader interpreted 276–300 screen-film mammograms, which were obtained under an institutional review board–approved, Health Insurance Portability and Accountability Act–compliant protocol. The need for informed consent was waived.

Radiologists read mammograms three times during a 20-month period (September, 2005 to May, 2007). Images were read in a mode we termed clinic–Breast Imaging Reporting and Data System (BI-RADS) (20), a mode rated under the ROC paradigm with an abnormality presence probability rating scale of 0–100, and a free-response ROC mode (21). The study was "mode balanced" in that three radiologists read mammograms with each of the modes first, three radiologists read each of the modes second, and three radiologists read each of the modes third (last) by using a block randomization scheme. The results of the clinic-BI-RADS mode are the focus of this article because it is most similar to clinical practice. In the future, we plan to report the results with the two other modes (ROC and free-response ROC) and their relationship to the results from the clinic-BI-RADS mode.

Four-view screen-film mammograms (ie, "current" mammograms) as well as a comparison mammogram (obtained at least 2 years before the study, when available, or 1 year before the study when it was the only available mammogram) as used during the original clinical interpretation were made available to the radiologists during the readings. Radiologists interpreted each mammogram as they would in the clinic and rated the right and left breast separately. The set read by each radiologist included a "common" set of 155 screen-film mammograms originally read in the clinic by other radiologists not participating in the study and a "reader-specific" set of mammograms that the reader had read clinically 2–6 years earlier. Common and reader-specific mammograms were mixed, and radiologists read all cases in one mode before moving to the next in the mode-balanced, case-randomized study that was managed by a comprehensive computer program. All ratings were recorded electronically and saved in a database.

Selection of Mammograms

The distribution of mammograms in the different categories was designed so that approximately 25% depicted positive findings (associated with verified cancers), approximately 10% depicted verified benign findings, and approximately two-thirds were either rated as

Advances in Knowledge

- Radiologists' performance in the clinical environment was significantly different than that in laboratory retrospective studies ($P = .035$).
- Interreader dispersion of the computed performance levels was significantly lower during the clinical interpretations than in the laboratory retrospective study ($P < .01$).

Implication for Patient Care

- Inferences regarding the performance of new technologies and clinical practices made as a result of retrospective laboratory observer studies may not always be valid or applicable.

Published online before print
10.1148/radiol.2491072025

Radiology 2008; 249:47–53

Abbreviations:
BI-RADS = Breast Imaging Reporting and Data System
CAD = computer-aided detection
ROC = receiver operating characteristic

Author contributions:
Guarantor of integrity of entire study, D.G.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, D.G.; clinical and experimental studies, C.S.C., C.M.H., L.A.H., M.A.G., R.L.P., W.R.P., R.S., J.H.S., L.P.W.; statistical analysis, A.I.B., H.E.R.; and manuscript editing, D.G., A.I.B., H.E.R.

negative during screening or recalled for a suspected abnormality but rated as negative during the diagnostic work-up that followed. Actually negative mammograms included *(a)* those originally given a BI-RADS rating of 1 or 2 (not recalled at screening) and verified as not showing cancer at least 1 year thereafter and *(b)* those originally given a BI-RADS rating of 0 and later found to be negative during a subsequent diagnostic work-up. Actually positive mammograms included *(a)* all available mammograms depicting pathologically confirmed cancers detected as a result of the diagnostic follow-up of a recall and *(b)* all false-negative mammograms—namely those mammograms that actually depict an abnormality that had been originally rated as negative (BI-RADS category 1) or benign (BI-RADS category 2) but later verified as positive for cancer within 1 year. Negative mammograms were selected in a manner that approximately one-third of the cases did not have a previous examination during the original interpretation to reflect our approximate clinical distribution. Therefore, 63% (635 of 1013) of the actually negative and 83% (293 of 354) of the actually positive mammograms had a previously obtained mammogram available for comparison.

Actually positive and negative mammograms were obtained from databases of the total screening population that are carefully maintained for quality assurance purposes and from our tumor registry. Actually negative mammograms were selected consecutively from our total screening population beginning with the first day of each calendar quarter from 2000 and continuing through 2003 (when the inclusion criteria and verification conditions were met) until the required predetermined number of cases in each category was reached. The time frame for searching for actually positive mammograms included all screening examinations performed between 2000 and 2004 to ensure the inclusion of as many consecutive screening-detected cancers by each of the nine participants as possible. Examinations were rejected if *(a)* any of the current or previous images were missing, *(b)* if "wires" marking scars of previous biopsies were visible or there was any indication (marking) of a palpable finding at the time of the screening that was placed during the examination (eg, BB) and, hence, was visible on the images, or *(c)* if a screening examination had been converted to a diagnostic examination during the same visit because of symptoms reported or discovered at screening. As a result of the selection protocol, a total of 354 "positive" mammograms (showing both screening-detected or missed but proved cancers) depicting the abnormalities in question were included in the study; 107 mammograms were rejected (45 with BB markings for palpable masses and 62 with wires marking scars and/or previous biopsies). The distributions of negative and positive mammograms with depicted abnormalities that were ultimately included in the study are summarized in Table 1. The average age of women whose mammograms were selected was 53.96 years (range, 32–93 years).

The use of computer-aided detection (CAD) was introduced into our clinical practice in mid-2001. Therefore, 752 of 1367 cases (55%) resulting from 671 of 1212 (55%) cases in the reader-specific set and 81 of 155 cases (52%) in the common set had been originally read in the clinic with CAD. We did not supply CAD results, however, because we previously determined that the effect of CAD on recall and detection rates in our practice, including with these very radiologists, was small (22), and CAD results had not been consistently kept throughout the period of interest.

Each mammogram was assigned a random identification number and cleaned; all identifying information, including time marks, was covered with black photographic tape. Study identification labels were affixed to all mammograms. Previously obtained images were identified and specifically marked with the number of months between the previous and current examination.

Study Performance

Observers were unaware of the specific aims of the study (ie, they were not told that they had previously read either all or some of the mammograms in the clinic) and received a general and a mode-specific "Instruction to Observers" document (23). The document included a general overview of the study setup and the process for reviewing and rating mammograms during a session and informed the reader that previous mammograms labeled with the approximate number of months between the relevant (current and previous) mammograms would be provided, if applicable. The document also described in detail how certain abnormalities (eg, asymmetric densities) should be scored and noted that the set of mammograms was enriched without any specific numbers or proportions. The instructions for the clinic-BI-RADS mode specifically stated that the reader was expected "to read and rate (interpret) the examinations as though they are being read in a

Table 1

Distribution of Mammograms with Specific Abnormalities Depicted

	Abnormality			
Mammogram Type	Mass	Microcalcifications	Both	Total
Negative*	0	0	0	492
Benign†	64	62	20	146
Recall‡	277	75	23	375
Positive§	181	139	34	354
Total	522	276	77	1367

* Rated BI-RADS 1 at screening and verified to be actually negative at follow-up.

† Rated BI-RADS 2 at screening and verified to be actually negative at follow-up.

‡ Rated BI-RADS 0 at screening and verified to be actually negative during diagnostic work-up.

§ All examinations depicting verified cancers.

screening environment.” The readers were not made aware of the specifics of each mode until the time they were scheduled to start that mode. A training and discussion session was implemented before the interpretations were begun. The training and discussion included a clear definition of abnormalities of interest and how these should be rated, as well as the protocol for using the computerized rating forms.

Mammograms to be read within each session included a randomized mix of images from the reader-specific and common sets. For example, as shown in Table 2, reader 1 read a total of 297 mammograms (155 common and 142 reader-specific) and reader 2 read 295 mammograms (155 common and 140 reader-specific). Note that each reader evaluated a different number of reader-specific mammograms because the set of mammograms included reader-specific images unique only to that particular reader. For each reading session, a randomized examination list was generated by a computer program that assigned a mammogram number to a specific slot number on the viewing alternator. All mammograms to be read during the specific session were loaded onto the film alternator according to the examination list generated by the computerized scheme. After matching the case number, observers reported their recommendations for each mammogram on a computerized scoring form. The number of mammograms interpreted during each reading session varied from 20 to 60, depending on what each participant’s schedule would allow and their own pace of reading; however, on average, about 15% of the mammograms were read per session. Answers could be changed while viewing an image until the “done” command was entered and final ratings were recorded.

During the clinic-BI-RADS mode, observers were first presented with a choice of rating the mammogram and each breast as “negative” (score, 1), “definitely benign” (score, 2), or “recommended for recall” (score, 0). If a benign or a recall rating was entered, observers were asked to identify the type of abnormality in question (ie, “mass,” “microcalcifications,” “other”) and could list more than one abnormality. If recall was recommended, a list of recommended follow-up procedures appeared and observers had to select at least one recommended procedure (eg, spot craniocaudal view/spot 90, spot craniocaudal view/whole breast, magnification craniocaudal view/90°, exaggerated craniocaudal view, tangential for calcifications, and/or ultrasonography).

Table 2

Performance Levels in Terms of Recall Rates of Actually Positive and Actually Negative Cases in the Clinic and during a Retrospective Laboratory Observer Performance Study

	Actual Clinical Data		Laboratory Rating Results	
Reader No.	Actually Negative	Actually Positive	Actually Negative	Actually Positive
	Reader-specific			
1	36/100 (0.360)	40/42 (0.952)	54/100 (0.540)	41/42 (0.976)
2	43/107 (0.402)	29/33 (0.879)	91/107 (0.850)	32/33 (0.970)
3	38/96 (0.396)	47/49 (0.959)	41/96 (0.427)	39/49 (0.796)
4	36/99 (0.364)	37/40 (0.925)	26/99 (0.263)	34/40 (0.850)
5	39/97 (0.402)	40/44 (0.909)	60/97 (0.619)	38/44 (0.864)
6	38/103 (0.369)	17/18 (0.944)	45/103 (0.437)	18/18 (1.00)
7	37/96 (0.385)	23/24 (0.958)	55/96 (0.573)	24/24 (1.00)
8	35/108 (0.324)	12/14 (0.857)	24/108 (0.222)	10/14 (0.714)
9	42/116 (0.362)	23/26 (0.885)	37/116 (0.319)	23/26 (0.885)
Total*	344/922 (0.374)	268/290 (0.919)	433/922 (0.472)	259/290 (0.895)
	Common			
Other†	31/91 (0.341)	58/64 (0.906)	. . .	. . .
1	. . .	. . .	50/91 (0.549)	58/64 (0.906)
2	. . .	. . .	74/91 (0.813)	60/64 (0.938)
3	. . .	. . .	35/91 (0.385)	49/64 (0.766)
4	. . .	. . .	21/91 (0.231)	47/64 (0.734)
5	. . .	. . .	41/91 (0.451)	56/64 (0.875)
6	. . .	. . .	39/91 (0.429)	49/64 (0.766)
7	. . .	. . .	60/91 (0.659)	60/64 (0.938)
8	. . .	. . .	24/91 (0.264)	37/64 (0.578)
9	. . .	. . .	30/91 (0.330)	53/64 (0.828)
Total	(0.341)	(0.906)	(0.457)	(0.814)

Note.—Data are given as numbers of mammograms rather than numbers of breasts.

* Simple average of reader-specific recall rates.

† Mammograms originally interpreted in the clinic by other radiologists not participating in the study.

Data Analyses

We focused our analysis on an examination-based rating, namely an examination in which only one breast containing cancer is treated as a “true positive” finding if a recall rating was given to either breast. For the purpose of this analysis, the primary sensitivity (or true-positive fraction) was estimated as a proportion of the “positive” mammograms out of all mammograms depicting verified cancers, and specificity (or 1 − false-positive fraction) was estimated as a proportion of the “negative” mammograms out of all verified “cancer-free” cases. In our primary analysis, we summarized performance over readers as a simple average.

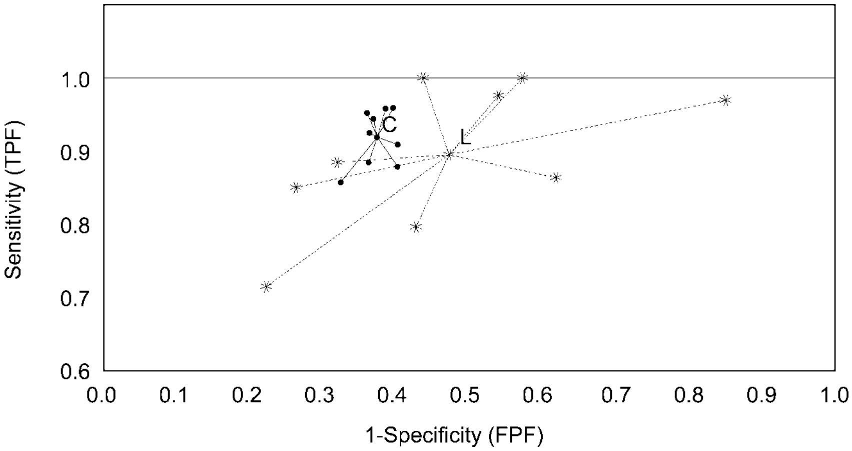
We conducted a statistical analysis that accounts for both the correlations between ratings on the same mammograms and for heterogeneity between observers’ levels of performance. The difference between performance levels in the clinic, namely the actual ratings (score of 0, 1, or 2) during the prospective clinical interpretation of each mammogram and the laboratory retrospective observer study was conducted by using hypotheses testing in the framework of a generalized linear mixed model with proc GLIMMIX SAS software (version 9.13; SAS Institute, Cary, NC).

We tested whether the average performance in the clinic and laboratory could be described with a single ROC curve. For this purpose, reader-specific and common sets of data were analyzed separately. We also verified the results of this analysis by performing an analysis conditional on the examinations with discordant ratings between the clinic and the laboratory.

In a separate analysis, we compared the dispersions (the average distance to the mean performance level) of the computed reader-specific operating characteristics. The comparison of performance dispersion was conducted only on reader-specific subsets by using Levene's test for paired data (24,25). We assessed the differences in spread of specificities, sensitivities, and distances from reader-specific to reader-averaged operating points.

Because a fraction of the actually benign mammograms should have led to a recall recommendation in the clinic regardless of the ultimate outcome—hence, affecting recall rates—we also analyzed the data after excluding the 425 mammograms (388 in the reader-specific and 37 in the common sets) with verified benign findings. In addition, because screening BI-RADS ratings were available for each breast separately, both from the actual clinical interpretations and the retrospective laboratory study, we also computed and compared the breast-based performance levels (sensitivity and specificity) and dispersion in performance levels among the nine radiologists. Namely, each breast (left and right) was considered a diagnostic unit, rather than a case-based analysis in which the most suspicious finding (hence, the corresponding rating) for either of the breasts was taken into account as the mammogram's final recommendation (or outcome).

We also assessed the possible effect, if any, of the use of CAD in 55% (671 of 1212) of cases during the original clinical interpretations of the reader-specific set on the results of our two primary analyses. Namely, the comparison of the average performance levels and the comparison of dispersions in performance levels among the nine radiologists. We estimated and compared the trend of the readers for performing on different ROC curves in the clinic and the laboratory for the two groups of cases initially evaluated with and without CAD. We computed the dispersions of reader-specific performance levels in the clinic and the laboratory for each of the groups of mammograms evaluated with and without CAD and verified the significance of the difference in dispersions adjusted for possible CAD effects. Last, we assessed whether or not there was an interaction between the possible effect of using CAD and the possible effect of the inclusion or exclusion of actually benign mammograms in the analysis.



Graph shows measured performance levels (sensitivity [true-positive fraction, *TPF*] and 1 − specificity [false-positive fraction, *FPF*]) and overall average performance (center points) for reader-specific sets of mammograms in the clinic (*C*) and laboratory (*L*). The differences in performance levels and dispersion (spread) among the nine radiologists in the clinic and in the laboratory are demonstrated.

Results

Table 2 provides the computed performance levels in the clinic and the laboratory for each of the nine radiologists. Both mean sensitivity and specificity were higher in the clinic than in the laboratory (sensitivity, 0.919 vs 0.895, respectively; specificity, 0.626 vs 0.528; Figure), although the levels for either sensitivity alone or specificity alone did not achieve statistical significance ($P >$ .1). This tendency was observed in both reader-specific and common sets. Four readers achieved higher sensitivity in the laboratory, albeit with a corresponding lower specificity.

Although the differences between sensitivity and specificity alone were not statistically significant, there were statistically significant differences between the clinic and laboratory for performing on different ROC curves due to the simultaneous decreases in the laboratory in both sensitivity and specificity levels (P = .035). The results of the unconditional analysis of the common sets of images agreed with the results for reader-specific sets (P = .027). A conditional model–based test on discordant mammograms only was significant ($P <$ .01), supporting the hypothesis that combined performance was higher in the clinic than in the laboratory.

There was a substantial difference in the spreads of the actual operating points in the clinic and laboratory on the reader-specific sets (Figure). The sample standard deviations of reader-specific specificities differed by a factor of 7.8 (0.0253 vs 0.1976), and the standard deviations of reader-specific sensitivities differed by a factor of 2.6 (0.0382 vs 0.0999). There was a significant difference ($P < .01$) between the dispersions (average distance to the mean performance levels) of the computed reader-specific operating points in the clinic and laboratory (0.0395 and 0.1870, respectively).

After benign mammograms were excluded, the differences between specificity levels in the clinic and laboratory increased as compared with the complete dataset, and the test for performing on different ROC curves was statistically significant ($P < .01$) for both reader-specific and common sets. The difference in spreads on reader-specific sets remained significant ($P < .01$) after the exclusion of actually benign mammograms.

The breast-based analyses demonstrated the same trend as the examination-based results. Both average sensitivity and average specificity levels in the clinic were higher than those in the laboratory (sensitivity, 0.901 vs 0.847, respectively; specificity, 0.792 vs 0.730). The sample standard deviations of reader-specific specificities differed by a factor of 4.3 (0.0254 vs 0.1096), and the standard deviations of reader-specific sensitivities differed by a factor of 1.8 (0.0473 vs 0.0867).

The use of CAD (or not) did not significantly ($P = .61$) affect the observation that performance levels in the clinic were superior to those in the laboratory. As related to the possible effect of the use of CAD on variability, the spread in performance levels (average distance from the mean performance level) in the clinic for the set interpreted with CAD was not smaller than that for the set interpreted without CAD (0.1252 and 0.0986, respectively). The ratios of performance dispersions between the clinic and the retrospective laboratory experiment were similar for the set of mammograms read in the clinic with CAD and the set of mammograms read in the clinic without CAD. The adjusted difference in dispersions of performance levels in the clinic and laboratory was statistically significant ($P = .025$). We note, however, that our study did not allow for an efficient, unbiased assessment of the possible effect of CAD on performance levels of individual readers or the dispersion in performance levels among readers as in studies when the same cases are read either prospectively or retrospectively with and without CAD by the same readers. There were no interactions ($P = .31$) between the effect of using CAD and the effect of including actually benign mammograms in that the inclusion or exclusion of mammograms depicting benign findings was similar whether CAD was used or not.

Discussion

Several retrospective studies demonstrated that radiologists' performance is relatively poor when interpreting screening mammograms and that radiologists' interreader variability is substantial (15–19,26). Inferences generated from these studies have been quoted numerous times and used as one of the primary reasons for the need for corrective measures (27). However, there are no substantial data about the "laboratory effect" or the correlation between performance in the clinic and laboratory experiments. This type of study is difficult to design because, in most areas, we do not have adequate quantifiable estimates of performance levels in the clinic. This is not the case in screening mammography, where the BI-RADS ratings can be used to estimate radiologists' performance levels in recalling (or not) women who ultimately are found to have breast cancer (or not). Hence, images from screening mammography were used in this study because the endpoint is typically binary (ie, recommendation to recall the women for additional work-up or not) and the outcome for most of those women who were not recalled can be verified with periodic follow-up.

The laboratory results in our study are similar to and consistent with those reported by Beam et al (16,19). However, on average, radiologists performed better in the clinic than in a retrospective laboratory experiment when interpreting the mammograms they themselves had read in the clinic. Reading "order effect," if any, would increase the observed differences in that all clinical readings were done first. These seemingly surprising results can be explained if one accepts that in the laboratory radiologists are aware that there is no effect on patient care; hence, the reporting pattern of at least some readers may change substantially. In addition, in the laboratory, radiologists are not affected by the pressure to reduce recommendations for recall per practice guidelines; hence, on average, their recall rate is higher (27). It is interesting that their average performance for mammograms they had actually read in the clinic was better than that for mammograms other radiologists had read in the clinic. Although this could be because they remembered some of the images, it is unlikely because, in addition to mixing these mammograms with others that they did not read, there was a long delay between the two readings (28) and they interpreted a very large number of images in between. It is quite possible that there is a "self-selection" bias; namely, if a radiologist is better at detecting certain types of depictions of cancers, then, over time, the type and distribution of cancers he or she detects is affected. Hence, when all cancers actually detected by a particular radiologist are used in a retrospective study, this set will be different than the type and distribution of cancers detected by other radiologists. Therefore, he or she will also be better at detecting "their own" type of cancers in a retrospective study. This finding suggests that continuous training (feedback) on cases missed by the individual radiologist rather that those missed by others may prove to be a better approach to continuing improvements in performance. Other clinical information (eg, patient or family history) may also have an effect on clinical decisions; however, it is not expected to be an important factor in the screening environment evaluated herein.

We note that we define "sensitivity" differently from other studies (29) in that here it is radiologists' sensitivity to actually depicted abnormalities. In addition, we do not have a full account of all false-negative findings and examinations "lost to the system" because some women with cancer may have relocated or decided to be treated elsewhere. These cases are not accounted for. Hence, our results are conditional on the dataset and readers in this study, and our conclusions must be independently validated. Our observations may be applicable solely to experienced radiologists who evaluate a high volume of mammograms (19).

Finally, the significantly higher performance in the clinic observed herein may have contributed to the difficulty in demonstrating actual significant improvements due to the use of CAD in some observational studies (22,30).

The mammograms used in this study were sampled in a manner that could have improved apparent estimates of sensitivity in the clinic because of the possible incomplete sampling of false-negative cases, making them potentially not representative of the true performance levels. However, we expect that on a relative scale the observed relationship between clinical and laboratory performance levels is similar in a true representative randomly selected sample set of examinations. This study may have implications on the clinical relevance of retrospective observer studies designed to assess and compare different technologies and practices.

In conclusion, when deciding whether to recall a woman for additional diagnostic examinations, experienced radiologists performed significantly better on average and, as important, more consistently in the clinic than in the laboratory when interpreting the same mammograms. Radiologists' interreader spread in performance levels was significantly lower during prospective clinical interpretations when the same clinical rating scale was used.

Acknowledgments: We thank Glenn Maitz, Jill King, Amy Klym, and Jennifer Stalder for their diligent and meticulous effort on this project.

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.
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.
3. 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(5):341–349.
4. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 2001;8(7):605–615.
5. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. Acad Radiol 2001;8(7):616–622.
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.
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.
8. Gur D, Rockette HE, Maitz GS, King JL, Klym AH, Bandos AI. Variability in observer performance studies experimental observations. Acad Radiol 2005;12(12):1527–1533.
9. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology 2003;228(1):10–14.
10. 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.
11. 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.
12. 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.
13. Egglin TK, Feinstein AR. Context bias: a problem in diagnostic radiology. JAMA 1996; 276(21):1752–1755.
14. Rutter CM, Taplin S. Assessing mammographers' accuracy: a comparison of clinical and test performance. J Clin Epidemiol 2000; 53(5):443–450.
15. 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.
16. 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.
17. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Womens Health 1998;7(4):443–449.
18. Esserman L, Cowley H, Eberle C, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 2002;94(5):369–375.
19. 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.
20. American College of Radiology. Breast imaging reporting and data system atlas (BI-RADS atlas). Reston, Va: American College of Radiology, 2003. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/BIRADSAtlas.aspx. Accessed October 1, 2003.
21. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol 2007;14(6):723–748.
22. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. J Natl Cancer Inst 2004;96(3):185–190.
23. Gur D, Rockette HE, Good WF, et al. Effect of observer instruction on ROC study of chest images. Invest Radiol 1990;25(3):230–234.
24. Grambsch PM. Simple robust tests for scale differences in paired data. Biometrika 1994; 81(2):359–372.
25. Levene H. Robust tests for equality of variances. In: Olkin I, ed. Contributions to probability and statistics. Palo Alto, Calif: Stanford University Press, 1960;278:–292.
26. Beresford MJ, Padhani AR, Taylor NJ, et al. Inter- and intra-observer variability in the evaluation of dynamic breast cancer MRI. J Magn Reson Imaging 2006;24(6):1316–1325.
27. Quality standards and certification requirements for mammography facilities—FDA. Interim rule with request for comments. Fed Regist 1993;58(243):67565–67572.
28. Hardesty LA, Ganott MA, Hakim CM, Cohen CS, Clearfield RJ, Gur D. "Memory effect" in observer performance studies of mammograms. Acad Radiol 2005;12(3):286–290.
29. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 2005;353(17):1773–1783.
30. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–1409.

Radiology 2008

This is your reprint order form or pro forma invoice

(Please keep a copy of this document for your records.)

Reprint order forms and purchase orders or prepayments must be received 72 hours after receipt of form either by mail or by fax at 410-820-9765. It is the policy of Cadmus Reprints to issue one invoice per order. **Please print clearly**.

Author Name ______________________________
Title of Article ______________________________
Issue of Journal__________________ Reprint # __________ Publication Date __________
Number of Pages__________________ KB # __________ Symbol Radiology
Color in Article? Yes / No (Please Circle)

Please include the journal name and reprint number or manuscript number on your purchase order or other correspondence.

Order and Shipping Information

Reprint Costs **(Please see page 2 of 2 for reprint costs/fees.)**

________ Number of reprints ordered $_________
________ Number of color reprints ordered $_________
________ Number of covers ordered $_________
Subtotal $_________
Taxes $_________

(*Add appropriate sales tax for Virginia, Maryland, Pennsylvania, and the District of Columbia or Canadian GST to the reprints if your order is to be shipped to these locations.*)

First address included, add $32 for
each additional shipping address $_________

TOTAL $_________

Shipping Address **(cannot ship to a P.O. Box) Please Print Clearly**

Name ______________________________
Institution ______________________________
Street ______________________________
City ______________ State _____ Zip __________
Country ______________________________
Quantity______________ Fax ______________
Phone: Day ______________ Evening ______________
E-mail Address ______________________________

Additional Shipping Address* **(cannot ship to a P.O. Box)**

Name ______________________________
Institution ______________________________
Street ______________________________
City ______________ State ______ Zip __________
Country ______________________________
Quantity ______________ Fax ______________
Phone: Day ______________ Evening ______________
E-mail Address ______________________________

* *Add $32 for each additional shipping address*

Payment and Credit Card Details

Enclosed: Personal Check ___________
Credit Card Payment Details _________

Checks must be paid in U.S. dollars and drawn on a U.S. Bank.

Credit Card: __ VISA __ Am. Exp. __ MasterCard
Card Number ______________________________
Expiration Date______________________________
Signature: ______________________________

Please send your order form and prepayment made payable to:

Cadmus Reprints
P.O. Box 751903
Charlotte, NC 28275-1903

Note: Do not send express packages to this location, PO Box.
FEIN #:541274108

Invoice or Credit Card Information

Invoice Address **Please Print Clearly**

Please complete Invoice address as it appears on credit card statement

Name ______________________________
Institution ______________________________
Department ______________________________
Street ______________________________
City ______________ State _____ Zip _______
Country ______________________________
Phone ______________ Fax ______________
E-mail Address ______________________________

Cadmus will process credit cards and *Cadmus Journal Services* will appear on the credit card statement.

If you don't mail your order form, you may fax it to 410-820-9765 with your credit card information.

Signature ______________________________ Date ______________________________

Signature is required. By signing this form, the author agrees to accept the responsibility for the payment of reprints and/or all charges described in this document.

Radiology 2008

Black and White Reprint Prices

Domestic (USA only)						
# of Pages	50	100	200	300	400	500
1-4	$221	$233	$268	$285	$303	$323
5-8	$355	$382	$432	$466	$510	$544
9-12	$466	$513	$595	$652	$714	$775
13-16	$576	$640	$749	$830	$912	$995
17-20	$694	$775	$906	$1,017	$1,117	$1,220
21-24	$809	$906	$1,071	$1,200	$1,321	$1,471
25-28	$928	$1,041	$1,242	$1,390	$1,544	$1,688
29-32	$1,042	$1,178	$1,403	$1,568	$1,751	$1,924
Covers	$97	$118	$215	$323	$442	$555

International (includes Canada and Mexico)						
# of Pages	50	100	200	300	400	500
1-4	$272	$283	$340	$397	$446	$506
5-8	$428	$455	$576	$675	$784	$884
9-12	$580	$626	$805	$964	$1,115	$1,278
13-16	$724	$786	$1,023	$1,232	$1,445	$1,652
17-20	$878	$958	$1,246	$1,520	$1,774	$2,030
21-24	$1,022	$1,119	$1,474	$1,795	$2,108	$2,426
25-28	$1,176	$1,291	$1,700	$2,070	$2,450	$2,813
29-32	$1,316	$1,452	$1,936	$2,355	$2,784	$3,209
Covers	$156	$176	$335	$525	$716	$905

Minimum order is 50 copies. For orders larger than 500 copies, please consult Cadmus Reprints at 800-407-9190.

Reprint Cover

Cover prices are listed above. The cover will include the publication title, article title, and author name in black.

Shipping

Shipping costs are included in the reprint prices. Domestic orders are shipped via UPS Ground service. Foreign orders are shipped via a proof of delivery air service.

Multiple Shipments

Orders can be shipped to more than one location. Please be aware that it will cost $32 for each additional location.

Delivery

Your order will be shipped within 2 weeks of the journal print date. Allow extra time for delivery.

Color Reprint Prices

Domestic (USA only)						
# of Pages	50	100	200	300	400	500
1-4	$223	$239	$352	$473	$597	$719
5-8	$349	$401	$601	$849	$1,099	$1,349
9-12	$486	$517	$852	$1,232	$1,609	$1,992
13-16	$615	$651	$1,105	$1,609	$2,117	$2,624
17-20	$759	$787	$1,357	$1,997	$2,626	$3,260
21-24	$897	$924	$1,611	$2,376	$3,135	$3,905
25-28	$1,033	$1,071	$1,873	$2,757	$3,650	$4,536
29-32	$1,175	$1,208	$2,122	$3,138	$4,162	$5,180
Covers	$97	$118	$215	$323	$442	$555

International (includes Canada and Mexico))						
# of Pages	50	100	200	300	400	500
1-4	$278	$290	$424	$586	$741	$904
5-8	$429	$472	$746	$1,058	$1,374	$1,690
9-12	$604	$629	$1,061	$1,545	$2,011	$2,494
13-16	$766	$797	$1,378	$2,013	$2,647	$3,280
17-20	$945	$972	$1,698	$2,499	$3,282	$4,069
21-24	$1,110	$1,139	$2,015	$2,970	$3,921	$4,873
25-28	$1,290	$1,321	$2,333	$3,437	$4,556	$5,661
29-32	$1,455	$1,482	$2,652	$3,924	$5,193	$6,462
Covers	$156	$176	$335	$525	$716	$905

Tax Due

Residents of Virginia, Maryland, Pennsylvania, and the District of Columbia are required to add the appropriate sales tax to each reprint order. For orders shipped to Canada, please add 7% Canadian GST unless exemption is claimed.

Ordering

Reprint order forms and purchase order or prepayment is required to process your order. Please reference journal name and reprint number or manuscript number on any correspondence. You may use the reverse side of this form as a proforma invoice. Please return your order form and prepayment to:

Cadmus Reprints
P.O. Box 751903
Charlotte, NC 28275-1903

Note: Do not send express packages to this location, PO Box.
FEIN #:541274108

Please direct all inquiries to:

Rose A. Baynard
800-407-9190 (toll free number)
410-819-3966 (direct number)
410-820-9765 (FAX number)
baynardr@cadmus.com (e-mail)

Reprint Order Forms and purchase order or prepayments must be received 72 hours after receipt of form.