the same effect and, conversely, reported that increasing the prevalence of pulmonary emboli in a test set of pulmonary arteriograms from 20% to 60% increased levels of suspicion, sensitivity, and area under the receiver operating characteristic curve. In a separate study, Gur et al (5) demonstrated that performance was higher and variability between observers was lower in clinical practice than in a laboratory test set, even when the same set of mammograms was read in both environments. Therefore, measures of variability between observers in the test situation may not accurately reflect variability in practice. Overall, the findings in these studies suggest that the level of enrichment of the case set and the reporting environment may be factors in achieving reasonably realistic conditions in research studies. I would be interested in the editorial authors' perspectives on how researchers should approach research design and reporting in light of this.

**Disclosures of Potential Conflicts of Interest: S.T.** No potential conflicts of interest to disclose.

## References


1. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology 2010;257(1):14–17.
2. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology 2003;228(1):10–14.
3. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: changing the confidence ratings. Acad Radiol 2007;14(1):49–53.
4. Egglin TKP, Feinstein AR. Context bias: a problem in diagnostic radiology. JAMA 1996; 276(21):1752–1755.
5. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology 2008;249(1):47–53.


## Response


From

Alexander A. Bankier, MD,* Deborah Levine, MD,* Elkan F. Halpern, PhD,† and Herbert Y. Kressel, MD*

*Radiology* Editorial Office, 800 Boylston St, 15th Floor, Boston, MA 02199* e-mail: *abankier@bidmc.harvard.edu*

Institute for Technology Assessment, Massachusetts General Hospital, Boston, Mass†


We thank Dr Taylor-Phillips for her interest in our recent editorial (1), "Consensus Interpretation in Imaging Research: Is There a Better Way?," and for her valuable comments. We agree with her comments in that a sufficiently high number of observers and thorough reporting of variability between observers are important, albeit not exclusive, aspects for reasonably realistic study conditions. Our editorial was focused solely on interobserver variability, not the broader issue of realistic study conditions that Dr Taylor-Phillips describes. The overall topic of study design for diagnostic performance studies cannot be effectively addressed in an editorial. Clearly, an important consideration for readers is how generally applicable the reported results are for a given study. For image interpretation, we believe an assessment of interobserver variability is an important component to help readers better understand the results that are being reported (2).

We appreciate Dr Taylor-Phillips' comments and recognize the need for more discussion on optimizing study design to facilitate the translation of diagnostic performance imaging studies to clinical practice.

**Disclosures of Potential Conflicts of Interest: A.A.B.** No potential conflicts of interest to disclose. **D.L.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: has received payment for expert witness testimony, visiting professor lectures, continuing medical education courses, editing books, and authoring book chapters; receives royalties from UpToDate and Amyrsis; received payment from American College of Radiology for development of the obstetrics ultrasonography online self-assessment module; will have stock options in SafeMed. Other relationships: none to disclose. **E.F.H.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: is a consultant for Hologic. Other relationships: none to disclose. **H.Y.K.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: receives royalties from Medrad for an endorectal coil. Other relationships: is a staff member at Beth Israel Deaconess Medical Center.

## References


1. Bankier AA, Levine D, Halpern EF, Kressel HY. Consensus interpretation in imaging research: is there a better way? Radiology 2010;257(1):14–17.
2. Levine D, Bankier AA, Halpern EF. Submissions to Radiology: our top 10 list of statistical errors. Radiology 2009;253(2):288–290.


## Limitations of Minimally Acceptable Interpretive Performance Criteria for Screening Mammography


From

Gregory P. Doyle, MBA,* Jay Onysko, MA,† Lisa Pogany, MSc,‡ Diane Major, PhD,‡ Judy Caines, MD,§ Rene Shumak, MD,‖ and Nancy Wadden, MD#

Breast Screening Program for Newfoundland and Labrador, 35 Major's Path, Suite 102, St John's, Newfoundland, Canada A1A 4Z9* e-mail: *gregory.doyle@easternhealth.ca*

Public Health Agency of Canada, Ottawa, Ontario, Canada†

National Institute of Public Health of Québec, Québec City, Québec, Canada‡

Nova Scotia Breast Screening Program, Halifax, Nova Scotia, Canada§

Ontario Breast Screening Program, Toronto, Ontario, Canada‖

Department of Diagnostic Imaging, St Clare's Mercy Hospital, St John's, Newfoundland, Canada#


## Editor:

The May 2010 *Radiology* article by Dr Carney and colleagues (1) piqued our interest. The authors set cut points to identify underperforming radiologists who might benefit from additional training. These include sensitivity less than 75%, specificity less than 88% or greater than 95%, recall rate (RR) less than 5% or greater than 12%, positive predictive value (PPV) less than 3% or greater than 8%, and cancer detection rate (CDR) less than 2.5 per 1000 interpretations.

It is difficult to see how these criteria could be used with confidence. The suggested cut points contain internal inconsistencies: PPV is mathematically derived (PPV = CDR/RR). Given the suggested minimum CDR (2.5 per 1000 interpretations) and the suggested maximum RR (12%), the lower bound for PPV would be 2.1%, which is outside the authors' acceptable range. Similarly, a CDR of 4.0 per 1000 interpretations would result from the authors' minimum acceptable RR (5%) and maximum PPV (8%).

A radiologist with a CDR of 5.0 per 1000 interpretations and an RR of 6% would have a PPV of 8.3%—too high. It is hard to see how additional training would benefit this radiologist. Otten et al (2) found that with an RR of greater than 5%, the CDR levels off, resulting in a disproportionate and undesirable rise in false-positive findings.

The authors note that certain combinations of outcomes will achieve an RR below the lower-bound, though these would not be problematic. This is difficult to reconcile, since high CDR and low RR always produce a high PPV. The authors implicitly acknowledge this, yet provide no concrete solutions. Thoughtful approaches for assessing the interrelationships between CDR, RR, and PPV have been published elsewhere (3).

Some of the normative data in the study comes from radiologists who had interpreted only 100 screening or diagnostic mammograms. Since reading 960 screening mammograms every 2 years is required for certification, the relevance of the resulting cut points can be questioned further.

It is not clear if the indicators are relevant for all patient populations. Factors such as the age of the screened population and screening history (first vs subsequent screening), not just for "high-risk populations," are intimately related to the performance of screening mammography.

**Disclosures of Potential Conflicts of Interest: G.P.D.** No potential conflicts of interest to disclose. **J.O.** No potential conflicts of interest to disclose. **L.P.** No potential conflicts of interest to disclose. **D.M.** No potential conflicts of interest to disclose. **J.C.** No potential conflicts of interest to disclose. **R.S.** No potential conflicts of interest to disclose. **N.W.** No potential conflicts of interest to disclose.

## References

1. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology 2010;255(2):354–361.
2. Otten JD, Karssemeijer N, Hendriks JH, et al. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. J Natl Cancer Inst 2005;97(10):748–754.
3. Blanks RG, Moss SM, Wallis MG. Monitoring and evaluating the UK National Health Service Breast Screening Programme: evaluating the variation in radiological performance between individual programmes using PPV-referral diagrams. J Med Screen 2001;8(1):24–28.

## Response

From
Patricia A. Carney, PhD,* Edward A. Sickles, MD,† Barbara S. Monsees, MD,‡ Lawrence W. Bassett, MD,§ and Diana L. Miglioretti, PhD‖
Department of Family Medicine and Department of Public Health and Preventive Medicine, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239-3098*
e-mail: *carneyp@ohsu.edu*
Department of Radiology, University of California San Francisco, San Francisco, Calif†
Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, Mo‡
Department of Radiology, University of California Los Angeles, Los Angeles, Calif§
Department of Biostatistics, University of Washington, Seattle, Wash‖

We greatly appreciate the comments of Dr Doyle and his colleagues. In addressing their first point (ie, difficulty in seeing how the criteria set in our study could be used with confidence because the cut points contain internal consistencies), we would like to point out that the cut point(s) for each individual metric were derived separately and were not intended to be so internally consistent that one bound of any given metric combined with a bound of another metric would always result in a within-bounds metric (1). To achieve such internal consistency would not only be an extremely complex endeavor, but would also result in such narrow bounds as to be unattainable by the majority of practicing U.S. radiologists. Rather, the cut points we derived were intended to serve as determinants of whether or not to perform a detailed review of the overall performance of a given radiologist, with the understanding that many radiologists so flagged would likely be determined to have acceptable overall performance.

Regarding the second point about our cut points being relevant only to radiologists practicing within the United States, we point out that the authors of the letter all are from Canada, a country in which screening mammography is centrally organized and provincially funded. Screening mammography in the United States is neither centrally organized nor fully government funded at the state or national level. Our metrics were derived by radiologists who practice only in the United States who were informed by normative data that come only from U.S. practices, a country in which lack of central organization precludes universal high-volume screening, perceived malpractice exposure likely results in much higher RRs than are observed elsewhere, and screening is performed more frequently (often annually) and for a wider range of patient ages (starting at age 40 years, with no upper age limit) than elsewhere.

Lastly, we agree that factors such as the age of the screened population and screening history (first vs subsequent screening), not just for high-risk populations, are intimately related to the performance of screening mammography, and we addressed this in the discussion section in our article.

**Disclosures of Potential Conflicts of Interest: P.A.C.** No potential conflicts of interest to disclose. **E.A.S.** No potential conflicts of interest to disclose. **B.S.M.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: expects less than $2000 for serving on the medical advisory board for Hologic, institution has a National Institutes of Health grant for photoacoustic breast imaging, received an honorarium from University of Alabama Birmingham for speaking. Other relationships: none to disclose. **L.W.B.** No potential conflicts of interest to disclose. **D.L.M.** Financial activities related to the present article:

Radiology

institution has grants from National Cancer Institute and American Cancer Society, institution has received travel support from American Cancer Society. Financial activities not related to the present article: institution has grants or grants pending from National Cancer Institute and American Cancer Society. Other relationships: none to disclose.

## Reference

1. Carney PA, Sickles EA, Monsees BS, et al. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology 2010;255(2):354–361.

## Errata

"Quantitative MR Imaging of Brain Iron: A Postmortem Validation Study." Radiology 2010;257(2):455–462

Page 459, right-hand column, line 3, the sentence should read: According to MR relaxation theory, R2* is made up of the sum of two rates, as follows: R2* = R2 + R2', where **R2'** is attributed to local magnetic field inhomogeneities and **R2** is associated with intrinsic tissue properties.

"Identifying Patients with Atypical Ductal Hyperplasia Diagnosed at Core-Needle Biopsy Who Are at Low Risk of Malignancy" [letter]. Radiology 2010; 257(3):893–894

Page 893, the fourth author's name should read Wei **Yang**, MD.