Radiology: Imaging Cancer

Test Sets and Real-Life Performance: Can One Predict the Other?

Denise Thigpen, MD • Jocelyn Rapelyea, MD

Denise Thigpen, MD, is the former director of breast imaging at Walter Reed National Military Medical Center and currently serves as assistant professor of radiology at The George Washington University School of Medicine. She is a practicing breast imager with a focus on teaching fellows, residents, and medical students, with special interests that include multimodality breast screening and supporting women in radiology.

Jocelyn Rapelyea, MD, is the vice chair of education and the radiology residency program director at The George Washington University Hospital. Dr Rapelyea specializes in diagnostic radiology and breast imaging and serves as the associate director of the breast imaging section. Dr Rapelyea has published a variety of abstracts, articles, and chapters and has made various national and international presentations.

A mammography screening program's primary intention is to diagnose subclinical breast cancer and reduce mortality and morbidity associated with advanced disease. The ongoing debate around initiation and screening intervals is due to various interpretations of mammography efficacy that must weigh against potential harms linked to recall rate and associated patient anxiety. There is a delicate balance between recall rate and cancer detection rate (CDR). Too low of a recall rate can increase the chances of missing a breast cancer, and too high of a recall rate leads to unnecessary additional examinations. In the United States, the Mammography Quality Standards Act requires periodic auditing to ascertain if a facility meets national mammography quality. In 2017, the U.S. Food and Drug Administration launched an initiative to improve mammography image quality, called the Enhancing Quality Using the Inspection Program, or EQUIP, to provide direct feedback to the technologist and improve image quality.

In an effort to augment these efforts, the American College of Radiology recommends supplemental evaluation of physician recall rate, CDR, and positive predictive value of recall (PPV1). A recent study by Rauscher et al (1) demonstrates that feedback and more robust outcome monitoring can improve performance metrics. Audits and similar tracking of individual physicians' metrics in the United Kingdom and many other European countries result in high breast CDRs while keeping recall rates low–an ideal combination for a screening examination. In the United Kingdom, participation in quality improvement programs leads to improved benchmark measures and can identify outliers that need remediation.

The Breast Screening Information System (BSIS) offers "real-life" individual reader performance data to improve practice habits. Although BSIS provides performance feedback, these practice considerations are recognized over a rolling 3-year period, as confirmation of normalcy (true-negative findings) would not be achievable until the woman presents herself for the next screening opportunity 3 years later. Attempts to achieve greater reader performance in a shorter interval have resulted in test set programs. Most U.K. readers participate in the PERFORMS (PERsonal PerFORmance in Mammographic Screening) self-assessment scheme, a free and confidential educational program established in 1991, which consists of testing with two enriched case sets each year. Offering the test set twice per annum allows for a fair estimate of overall ability and reduces the likelihood of outside variables that may affect a reader's performance on a single test set (2). Similarly, Australia and New Zealand offer a test set program called BREAST (BreastScreen Reader Assessment Strategy), which offers some correlation with real-life performance (3). Test sets can provide immediate feedback on cases with pathologically proven or long-term confirmation of disease. Whether test set data are generalizable to real-life practice is a question of great interest.

In this issue of *Radiology: Imaging Cancer*, Chen et al set out to demonstrate the relationship between real-life performance and test set performance for readers in the National Health Service Breast Screening Programme (NHSBSP) (4). While the authors acknowledge that reading behavior in test conditions may not reflect real-life practice patterns, given the known "laboratory effects" inherent in a nonclinical setting, they do show positive correlations between PERFORMS and real-life BSIS data. Out of the 706 readers invited, 582 NHSBSP readers consented to participate in this study. Their real-life BSIS data were obtained over a 3-year period (2013–2016) and compared with their PERFORMS sets during that same time interval (at least five sets of cases required). A total of 452 readers met all inclusion criteria, and their CDR, recall rate, and PPV were compared between the two data sets. Additionally, any reader who was flagged as an "outlier" on any single PERFORMS set was also compared with the

From the Department of Radiology, The George Washington University Medical Center, 2150 Pennsylvania Ave NW, Washington, DC 20037. Received September 4, 2020; revision requested September 4; revision received September 4; accepted September 8. **Address correspondence to** J.R. (e-mail: *jrapelyea@mfa.gwu.edu*).

See also the article by Chen et al in this issue. Conflicts of interest are listed at the end of this article.

 • ©RSNA, 2020

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

"nonoutliers." An outlier is a term defined by PERFORMS as a reader "whose test performance falls more than one and a half times the inter-quartile range below the 25th percentile in terms of either cancer detection rate or the area under the curve of the receiver operating characteristic analysis (or both)" (4).

The mean CDR from real-life BSIS data were 7.79 per 1000 women screened (0.78%), which did show a statistically significant positive correlation with the CDR in the PERFORMS test sets of 22.86%. Additionally, the recall rates also showed a statistically significant positive correlation between the two sets, with the mean real-life recall rate of 5.29% compared with PERFORMS mean recall rate of 37.49%. The PPV, a reflection of both the CDR and recall rate, also demonstrated a significant positive correlation between the two data sets. In general terms, this means that a reader with a higher CDR on the PERFORMS test sets will *tend* to have a higher CDR in real life, and the same is true for the other variables analyzed. However, pooled data can only describe *tendencies* and cannot directly correlate performance on the individual level. A comparison of PERFORMS outliers to nonoutliers demonstrated that outliers have significantly lower performance in real life than their peers in CDR and PPV but did not differ significantly in recall rate. The authors recognize that nearly 20% of PERFORMS readers did not consent to this study, which contributes to selection bias that has not been accounted for.

Germane to this discussion is a comparison of benchmarks between centralized programs such as NHSBSP and decentralized systems such as in the United States. For comparison, mean benchmark CDRs in the United States have been reported as 5.1 cancers per 1000 women screened, mean callback rate of 11.6%, and PPV1 of 4.4% (5). Understanding the disparities between the data requires recognizing the many differences in the landscapes in which these programs occur. Screening in the United States occurs in a wide range of settings, including academic centers, private practices, and health maintenance organizations, as opposed to the centralized national organization of the United Kingdom's program. The populations screened are also quite different: In the United States, women are advised to screen annually beginning at age 40, while in the United Kingdom, women are only screened every 3 years beginning at age 50. Radiologists in the United States have different reading requirements, which is, on average, five to seven times fewer than their U.K. counterparts (6). Double reading, while standard in the United Kingdom, is far less common in the United States. Interventions attempting to decrease recall rates while maintaining CDRs have been studied in the United States. Mullen et al required radiologists to review their own recalled cases (including diagnostic evaluation outcomes and biopsy results) and found recall rates dropped significantly. The addition of a consensus discussion improved metrics further while improving CDR (7).

It would be interesting to note if any other trends could be elucidated from the data presented by Chen et al in this issue. For example, if the outlier status was applied at the beginning of the 3-year period, could this result in individual improvement over time, as a result of remediation or increased experience? Do readers who are consistently flagged as outliers on multiple PERFORMS sets show a stronger correlation to real-life performance, as opposed to those who only demonstrate outlier status on a single test set? Further evaluation in these areas could make this data more useful to an individual reader. Test set results can be especially important to guide low-volume readers, early-career readers, and others who may not have accrued sufficient real-life feedback to refine their reading practices. Test sets can also provide invaluable input on potential areas of weakness to allow for targeted supplemental training or remediation. This study adds to the understanding of the relationship between test set performance and real-life performance, with the ultimate goal of achieving the delicate balance of reasonable recall rates and excellent CDRs.

Disclosures of Conflicts of Interest: D.T. disclosed no relevant relationships. **J.R.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed payment received from GE Healthcare for lectures including service on speakers bureaus. Other relationships: disclosed no relevant relationships.

References

1. Rauscher GH, Tossas-Milligan K, Macarol T, Grabler PM, Murphy AM. Trends in Attaining Mammography Quality Benchmarks With Repeated Participation in a Quality Measurement Program: Going Beyond the Mammography Quality Standards Act to Address Breast Cancer Disparities. J Am Coll Radiol 2020. 10.1016/j.jacr.2020.07.019. Published online August 7, 2020.
2. Gale AG. Performs: A self-assessment scheme for radiologists in breast screening. Semin Breast Dis 2003;6(3):148–152.
3. Soh BP, Lee WB, Mello-Thoms C, et al. Certain performance values arising from mammographic test set readings correlate well with clinical audit. J Med Imaging Radiat Oncol 2015;59(4):403–410.
4. Chen Y, James JJ, Cornford EJ, Jenkins J. The Relationship between Mammography Readers' Real-Life Performance and Performance in a Test Set–based Assessment Scheme in a National Breast Screening Program. Radiol Imaging Cancer 2020;2(5):e200016.
5. Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. Radiology 2017;283(1):49–58.
6. Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United kingdom. JAMA 2003;290(16):2129–2137.
7. Mullen LA, Panigrahi B, Hollada J, Panigrahi B, Falomo ET, Harvey SC. Strategies for Decreasing Screening Mammography Recall Rates While Maintaining Performance Metrics. Acad Radiol 2017;24(12):1556–1560.