Evaluating the Impact of New Imaging Tests: Promises and Pitfalls

Daniel A. Pryma, Angela DeMichele, David A. Mankoff

Correspondence to: Daniel A. Pryma, MD, Department of Radiology, University of Pennsylvania Perelman School of Medicine, 3400 Spruce St, Donner 110-C, Philadelphia, PA 19104 (e-mail: daniel.pryma@uphs.upenn.edu).

The introduction of a new diagnostic imaging test into cancer clinical practice creates both anticipation and confusion. The use of 18F-fluorodeoxyglucose (18FDG) positron emission tomography (PET)/computed tomography (CT) in breast cancer is a good example. Early studies indicated high sensitivity for regional nodal metastases (1), whereas later studies in patient populations more representative of early breast cancer diagnosis showed much poorer performance (2). More recently, data have supported the potential utility of 18FDG PET/CT in locally advanced breast cancer (3,4); based on these data, current clinical guidelines recommend 18FDG PET/CT as an optional test in stage III breast cancer (5).

Nevertheless, the data in support of the use of 18FDG PET/CT for locally advanced breast cancer are from relatively small and largely retrospective series, and controversy remains about the utility of 18FDG PET/CT in the initial evaluation of patients with newly diagnosed breast cancer at high risk for recurrence.

In this issue of the Journal, Groheux et al. (6) shed some additional light on this topic. In one of the largest series to date, the authors describe their prospective evaluation of 18FDG-PET/CT in consecutive patients over the course of 5 years with newly diagnosed stage II or III breast cancer. Restricting inclusion in their series to women without positive sentinel lymph nodes, they found that 18FDG PET/CT detected N3 disease in 16% of patients overall and was positive for suspected M1 disease in 17% to 47% of patients with stage IIB to stage IIIC breast cancer. Importantly, in the subset of patients with long-term follow up, 18FDG PET/CT evidence of M1 disease (pathologically confirmed in 40% of patients) was predictive of disease-specific survival. The Groheux et al. study (6) was a long-term undertaking and supports the utility of 18FDG-PET/CT in locally advanced breast cancer and possibly in stage IIB and IIIA disease.

Although the Groheux et al. study (6) adds considerably to the data supporting the use of FDG PET/CT in stage IIB to stage IIIIC disease, the study also illustrates some of the difficulties in thoroughly investigating the utility and prognostic value of diagnostic tests. Clear guidance on optimal evaluation and reporting of diagnostic tests has been published (7–9). These guidelines form an appropriate and robust directive for diagnostic testing and have become the standard for diagnostic laboratory assays, but they are difficult, and in some cases impossible, to fully satisfy for diagnostic imaging studies. In the early evaluation of a new diagnostic imaging modality, it is common to collect data prospectively and then determine whether results correlate with endpoints of interest in a post hoc analysis. Although this data-mining approach is appropriate in hypothesis generating and for secondary endpoints, it is critical for primary evaluation of prognostic biomarkers that the putative diagnostic criteria be prospectively defined and that the study be powered to test these variables. Retrospective definition of diagnostic criteria that are supported by the data introduces a high likelihood of bias in favor of the experimental test, whereas an underpowered study can yield incomplete or inaccurate results.

For laboratory-based diagnostic tests, conducting follow-up on studies with prospectively determined diagnostic criteria can be carried out ethically and with relatively modest cost. However, for an imaging study, costs, logistics, and potential medical risks make large-scale, prospective, phase III studies a challenge. For example, whereas a study of a laboratory test often requires little direct subject participation other than perhaps the removal of an additional tube of blood during routine phlebotomy or consent for use of archival pathologic source material, imaging trials require significant subject time commitment and may involve the need for an invasive biopsy to confirm results, adding risk and expense. Whereas enrollment in drug trials is frequently brisk when there is the potential for clinical benefit from the experimental therapeutic, blinded diagnostic imaging trials often struggle to meet accrual goals because the patient is being asked to participate with no possibility of direct benefit. Thus, because of a variety of factors, but to a great degree because of cost and feasibility, many evaluations of diagnostic imaging tests are undertaken with the experimental test being performed as part of standard of care, as was done in this case.

Can a new diagnostic test be accurately assessed in the context of clinical care? Although such studies may be prospective, they are not randomized, and there is often an incomplete reference standard. There is bias introduced because only those patients who choose to undergo the “experimental” test as part of their standard of care are eligible for the study. Most important, in the absence of blinding, the “experimental” imaging test results are provided to patients and providers, who are often compelled to use the results of the experimental test to direct decisions on whether to use tests that are part of the reference standard (eg, whether or not to get another imaging test or a biopsy to confirm the results of the modality under study) or to direct therapy. In the Groheux et al. article (6), for example, standard imaging was specifically chosen in many cases to clarify the 18FDG-PET/CT findings rather than act as a standard companion to it. Furthermore, the experimental test could be used to direct therapy, whereas the ideal study protocol would prohibit disclosure of results of the experimental test to prevent alteration of subsequent management. However, blinding...
clinicians to the results of a test already in clinical use may create ethical challenges and will certainly limit interest on the part of both patients and clinicians to participate. Finally, blinding is impossible if the test in question is being billed and reimbursed as part of standard of care.

To calculate sensitivity, specificity, and likelihood ratios for an experimental diagnostic test, all positive and negative test findings must be adjudicated by the reference standard, but in clinical trials of imaging modalities, this is rarely possible. For example, in the case of the internal mammary node, confirmatory biopsy may be unsafe or technically infeasible. In many studies, the evolution of a lesion on follow-up is used to subsequently validate the test findings, but given the dynamic nature of cancer and cancer therapy, this approach may be misleading because interventions that impact the findings of the test may have occurred in the interim. Given the lack of a consistent, independent reference standard and a blinded approach or control group, it can be difficult to judge the prognostic impact of $^{18}$FDG-PET/CT using data gleaned from emerging clinical practice.

The Groheux et al. study (6) is important in shedding light on the potential use of $^{18}$FDG-PET/CT in stage II and III breast cancer to accurately stage patients, a relevant clinical dilemma. Their results provide supportive evidence for a role for $^{18}$FDG-PET/CT in determining stage and prognosis of high-risk patients. However, the definitive evaluation of $^{18}$FDG-PET/CT as a prognostic biomarker in stage II and III breast cancer would require a confirmatory study with strict patient selection rules, independent reference standards, and blinding to imaging results or a control group. Such a study would be challenging, and likely expensive, but it would generate the important data needed to guide the use of a powerful but costly diagnostic imaging modality in a highly prevalent disease. This has been accomplished for $^{18}$FDG-PET/CT for early-stage breast cancer, for which limited use was confirmed (10). Perhaps it is time to conduct a similar study to direct the appropriate use of what appears to be a useful test for higher-stage breast cancer.

**References**


**Notes**

The authors declare no relevant conflicts of interest.

**Affiliations of authors:** Department of Radiology (DAP, DAM) and Department of Medicine, Biostatistics and Epidemiology (AD), University of Pennsylvania Perelman School of Medicine, Philadelphia.

In the article “International Multicenter Tool to Predict the Risk of Nonsentinel Node Metastases in Breast Cancer,” Meretoja et al. (1) described a predictive tool that could be used to assess the risk of additional axillary metastases after a positive sentinel node biopsy. This model, which incorporated data from five centers and 1000 patients, has undergone extensive internal and external validation in identifying nine statistically significant factors for positive nonsentinel nodes.

Many questions about the management of the axilla in breast cancer, particularly questions about the role of axillary node dissection in the setting of a positive sentinel node, remain. Much of the debate centers around the still-uncertain impact of axillary node dissection on the incidence of regional recurrence, survival, and morbidity, particularly because the majority of patients with a positive sentinel node will not have additional positive axillary nodes found on completion of axillary node dissection.

**Does Predicting Positive Nonsentinel Nodes Answer the Question of Axillary Dissection and Provide a Benefit?**

Carol Slomski, Thomas B. Julian

Correspondence to: Thomas B. Julian, MD, National Surgical Adjuvant Breast and Bowel Project, 320 E North Ave, Pittsburgh, PA 15212 (e-mail: tjuan@nasbp.org).