A Flexible, Multifaceted Approach Is Needed in Health Technology Assessment of PET

PET is finding numerous clinical applications, particularly in oncology, in which it is used for staging, treatment planning, response assessment, prognostication, and recurrence monitoring, but also in cardiology, neurology, and infectious disease (*1,2*). We now have several systematic reviews and decision and economic analyses of PET in various clinical contexts, as well as methodologic appraisals of randomized controlled trials (RCTs) in which PET was used as a comparator (*3–5*). Yet, the widespread adoption and reimbursement of PET has proven controversial. In addition to the high cost of the technology, there are frequent disagreements about the strength of the evidence supporting its various applications (*6*).

In this issue of *The Journal of Nuclear Medicine,* Siepe et al. report an up-to-date methodologic appraisal of 14 published and 15 planned RCTs of PET (*7*). They found that trials typically used comparators that reflected current clinical practice (at the time of conduct) but often did not provide evidence on patient-important outcomes and had sample sizes that were too small to allow for the reliable detection of differences in such outcomes. Readers could benefit from additional information about the clinical context and the intended use (e.g., clinical setting and specific management context) and role (e.g., as a replacement, add on, triage) of PET in the included trials, but such information is often poorly reported in the primary studies.

See page 1228

Siepe et al. (*7*) provide a thoughtful summary of randomized evidence on PET; however, it is important for readers to keep in mind that not all aspects of the evaluation of PET, or indeed any medical test, require randomized studies. In this invited perspective, we highlight challenges in comparative effectiveness research on medical tests as they apply to PET and argue that methodologic pluralism—a flexible, multifaceted approach to the choice of research methods—is more appropriate for the evaluation of this technology.

FUNDAMENTAL CHALLENGES IN THE EVALUATION OF MEDICAL TESTS

The assessment of medical tests poses 2 fundamental challenges for clinical researchers. First, the impact of tests on patient-centered outcomes is primarily indirect. Tests provide information that may modify patients' and physicians' thinking and can influence subsequent diagnostic and therapeutic decisions. Thus, the impact of tests on patient outcomes is typically mediated by the intervening medical care. Importantly, testing has an effect only among the subset of patients for whom test results influence management and, of those, only among the subset for whom changes in management impact outcomes (*8*). The magnitude of the effect of a test depends crucially on the effectiveness of subsequent care strategies.

Second, when evaluating medical tests, researchers must consider multiple variants of the technology of interest—a situation that is not usually encountered in studies of well-defined therapeutic interventions. PET exemplifies this problem because of the large number of tracers, image acquisition techniques, and qualitative and quantitative assessment methods that define alternative imaging protocols. Additional variability is introduced by differences in image interpretation among providers and the rapid evolution of PET technology over time. Further, decision makers are generally interested in comparisons of the various PET-based strategies against all feasible alternative testing strategies, not only against not using PET. As technology advances, an increasing number of PET-based and non–PET-based technologies (e.g., PET combined with MR imaging and other morphologic imaging modalities) will become available, and their comparative effectiveness will have to be evaluated.

These fundamental challenges have important implications for the assessment of PET. First, it should be clear that comprehensive studies comparing alternative testing strategies with respect to test performance, impact on patient management, and patient-relevant outcomes in a single population of interest are almost never available. Thus, synthesis methods are needed to piece together the evidence puzzle comprising separate studies of the impact of various tests and test-directed treatments on outcomes across heterogeneous populations. When such diverse sources of evidence are considered, global subjective judgments on whether a test works are often inadequate, and formal methods for evidence synthesis are needed (*9*).

Second, the availability of many PET variants, the multitude of clinical contexts in which PET can be used, and the large number of competing technologies mean that RCTs addressing a specific clinical use of PET will by necessity be scarce. The cost of conducting trials comparing all tests of interest in every pertinent clinical setting is simply prohibitive. Because PET, like any test used to inform patient management, affects outcomes only on a subset of enrolled participants (those for whom it suggests a change in management and for whom that change affects outcomes), designing and conducting adequately powered studies—particularly RCTs—is bound to be difficult. Thus, it is unrealistic to expect that decisions about the adoption or coverage of PET should be exclusively based on direct randomized comparisons. Further, not making full use of the large number of nonrandomized studies of PET leads to an ever-expanding list of topics for which evidence is labeled insufficient for decision making.

Received Jun. 17, 2014; revision accepted Jun. 19, 2014.

For correspondence or reprints contact: Issa J. Dahabreh, Center for Evidence-Based Medicine, Brown University, P.O. Box G-S121-8; Providence, RI 02912.

E-mail: issa_dahabreh@brown.edu

Published online Jul. 21, 2014.

DOI: 10.2967/jnumed.114.142331

THE NEED FOR METHODOLOGIC PLURALISM IN THE ASSESSMENT OF PET

To overcome the challenges in assessing the clinical value of PET, we need a flexible, multifaceted approach for efficient evidence generation and synthesis. The many proposals for a phased assessment of medical tests, using hierarchic schemes that encompass technical validity, clinical validity, clinical utility, and economic or societal impact, are important for addressing this need (*10*). In broad terms, evidence is needed on the diagnostic or predictive accuracy of PET in specific clinical contexts; on the impact of the information from PET on subsequent diagnostic and therapeutic decisions; and on the impact of the use of PET on patient outcomes, including morbidity, mortality, and quality of life and functioning (*11*). The generation of such evidence requires cascades of studies, each with its own design and analysis requirements. Operationalizing this research agenda requires embracing methodologic pluralism as a necessary aspect of the evaluation process (*12*).

Broader and higher-quality nonrandomized studies of PET should be conducted, and new methods for their synthesis need to be developed. Despite the availability of a large number of cohort studies on the test performance of PET in various clinical settings, more studies are needed that compare test performance among alternative PET-based strategies and competing modalities (e.g., using paired designs with multiple tests applied in the same group of patients) (*13*). Future studies of test performance should aim to minimize bias (*14–16*) and report findings in a way that facilitates synthesis (*17*). Methodologic advances could lead to more informative evidence synthesis of PET studies. Most test performance meta-analyses to date are focused on a single test modality and use relatively simple statistical methods that do not take full advantage of the available data (*4*). New methods, along with better reporting in individual studies, would allow meta-analyses to provide estimates of comparative test performance (*18*). More importantly, the synthesis of studies of test performance should be viewed as the first step toward modeling assessments of the impact of testing. These assessments can provide quantitative summaries of the evidence and projections useful for clinical decision making and can also be used to plan future research.

Often perceived as the last step in test evaluation (in the form of decision or cost-effectiveness analysis), modeling can instead serve multiple roles throughout the process (*19*). Mathematical models can be used to communicate assumptions; summarize and extrapolate empirical findings; design future studies; and incorporate information on costs, patient preferences, and clinical outcomes. For example, modeling is the most suitable approach for deciding if PET is accurate enough in a particular clinical context to warrant its inclusion in clinical care strategies and in adaptive clinical trials. Modeling efforts can be substantially strengthened by the conduct of studies assessing the impact of testing on intermediate outcomes, such as the impact of tests on diagnostic or therapeutic decisions (*20,21*). These studies provide a critical link in the chain of evidence between test performance and clinical outcomes. Combined with evidence on the effectiveness of specific treatments, they can be used to quantify the impact of tests on outcomes, to identify tests that do not require further evaluation (e.g., if their impact on therapeutic decisions is negligible), and to inform the design of future comparative studies (e.g., by facilitating realistic power calculations) (*22*). Modeling is also helpful when data on benefits and harms of testing need to be considered jointly with economic costs and patient preferences (values).

Tests that appear most promising can be evaluated further using routinely collected data. Observational studies of alternative test strategies—especially when based on large databases of previously collected data (e.g., claims records, registries)—can be used to study large samples of patients who are representative of those seen in clinical practice (*23*). Such studies can be used to assess comparative test performance, but more importantly, they can be used to assess the impact of tests on clinical outcomes and costs. Currently available data sources are, in many cases, inadequate for the assessment of medical tests; however, richer datasets (e.g., incorporating data from electronic health records) are becoming available and hold substantial promise in the assessment of tests.

In some cases, the evidence from test performance studies, observational comparative studies, and modeling will be enough to support the use of tests in clinical practice. Occasionally, the decision can be based on heuristic rules (*24*); but, in most cases, mathematical modeling will be needed to combine information across studies and perform sensitivity analyses to assess the impact of modeling assumptions (*25*). When evidence is deemed insufficient, RCTs may be required. In general, simple parallel group RCTs are unlikely to be the most efficient design (*26*). For example, paired designs, in which all patients receive 2 (or more) tests, but only those with discrepant results are randomized to alternative treatments, are more efficient (*27*). A common limitation of RCTs is that they cannot provide information on all relevant patient populations, either because the target populations are too small or because the trials themselves use restrictive selection criteria. Novel methods and observational data can be leveraged to assess the representatives of RCT populations and to extrapolate trial results (*28,29*).

In summary, state-of-the-science evaluation of PET should use a broad range of methods. RCTs, when available, are likely to be only a small part of the evidence base for PET and should be understood to represent only one of the many possible inputs required for comprehensive evaluation. When available, evidence from RCTs should be synthesized with that from other sources, underscoring the broad utility of mathematical modeling in the evaluation of tests. Modeling can provide a unifying framework for the efficient generation and synthesis of evidence to inform clinical and policy decisions regarding PET.

Issa J. Dahabreh
Constantine Gatsonis
Brown University
Providence, Rhode Island

REFERENCES

1. Fletcher JW, Djulbegovic B, Soares HP, et al. Recommendations on the use of ^{18}F-FDG PET in oncology. *J Nucl Med.* 2008;49:480–508.
2. Mittra E, Quon A. Positron emission tomography/computed tomography: the current technology and applications. *Radiol Clin North Am.* 2009;47:147–160.
3. Facey K, Bradbury I, Laking G, Payne E. Overview of the clinical effectiveness of positron emission tomography imaging in selected cancers. *Health Technol Assess.* 2007;11:iii–iv, xi–267.
4. Dahabreh IJ, Chung M, Kitsios GD, et al. Survey of the methods and reporting practices in published meta-analyses of test performance: 1987 to 2009. *Res Synth Methods.* 2013;4:242–255.
5. Scheibler F, Zumbe P, Janssen I, et al. Randomized controlled trials on PET: a systematic review of topics, design, and quality. *J Nucl Med.* 2012;53:1016–1025.

6. Ware RE, Hicks RJ. Doing more harm than good? Do systematic reviews of PET by health technology assessment agencies provide an appraisal of the evidence that is closer to the truth than the primary data supporting its use? *J Nucl Med.* 2011;52(suppl 2):64S–73S.
7. Siepe B, Hoilund-Carlsen PF, Gerke O, Weber WA, Motschall E, Vach W. The move from accuracy studies to randomized trials in PET: current status and future directions. *J Nucl Med.* 2014;55:1228–1234.
8. Langlotz CP. Overcoming barriers to outcomes research on imaging: lessons from an abstract decision model. *Acad Radiol.* 1999;6(suppl 1):S29–S34.
9. Eddy DM. Practice policies: where do they come from? *JAMA.* 1990;263:1265–1275.
10. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. *Med Decis Making.* 2009;29:E13–E21.
11. Gatsonis C. The promise and realities of comparative effectiveness research. *Stat Med.* 2010;29:1977–1981.
12. Tunis SR, Benner J, McClellan M. Comparative effectiveness research: policy context, methods development and research infrastructure. *Stat Med.* 2010;29: 1963–1976.
13. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med.* 2013;158: 544–554.
14. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med.* 1987;6:411–423.
15. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282:1061–1066.
16. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ.* 2006;174: 469–476.
17. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med.* 2003;138:40–44.
18. Trikalinos TA, Hoaglin DC, Small KM, Schmid CH. *Evaluating Practices and Developing Tools for Comparative Effectiveness Reviews of Diagnostic Test Accuracy: Methods for the Joint Meta-Analysis of Multiple Tests*. Rockville, MD: Agency for Healthcare Research and Quality; 2013.
19. Trikalinos TA, Kulasingam S, Lawrence WF. Chapter 10: deciding whether to complement a systematic review of medical tests with decision modeling. *J Gen Intern Med.* 2012;27(suppl 1):S76–S82.
20. Hillner BE, Siegel BA, Liu D, et al. Impact of positron emission tomography/computed tomography and positron emission tomography (PET) alone on expected management of patients with cancer: initial results from the National Oncologic PET Registry. *J Clin Oncol.* 2008;26:2155–2161.
21. Hillner BE, Siegel BA, Shields AF, et al. The impact of positron emission tomography (PET) on expected management during cancer treatment: findings of the National Oncologic PET Registry. *Cancer.* 2009;115:410–418.
22. Hinchliffe SR, Crowther MJ, Phillips RS, Sutton AJ. Using meta-analysis to inform the design of subsequent studies of diagnostic test accuracy. *Res Synth Methods.* 2013;4:156–168.
23. Hillner BE, Liu D, Coleman RE, et al. The National Oncologic PET Registry (NOPR): design and analysis plan. *J Nucl Med.* 2007;48:1901–1908.
24. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med.* 2006;144:850–855.
25. Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. *Acad Radiol.* 2000;7:681–683.
26. Vach W, Hoilund-Carlsen PF, Gerke O, Weber WA. Generating evidence for clinical benefit of PET/CT in diagnosing cancer patients. *J Nucl Med.* 2011;52 (suppl 2):77S–85S.
27. Lu B, Gatsonis C. Efficiency of study designs in diagnostic randomized clinical trials. *Stat Med.* 2013;32:1451–1466.
28. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A Stat Soc.* 2001;174:369–386.
29. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol.* 2010;172:107–115.