# Is There Evidence for Evidence-Based Medical Imaging?

ΓΝΩΘΙ ΣΑΥΤΟΝ – *Know thyself*
Inscription at the Apollo temple in Delphi (*1*)

In this supplement to *The Journal of Nuclear Medicine,* Ware and Hicks provide scathing criticism of the misuse of evidence-based medicine in health technology assessments (*2*). Their critique focuses on Australian health technology assessments on the use of PET in oncology, but similarly controversial health technology assessments on PET have been performed in other countries as well. In Germany, the Institute for Quality and Cost Effectiveness in Health Care (IQWIG) has recently concluded that there is no evidence for the use of PET/CT in malignant lymphomas, squamous cell carcinomas of the head and neck, ovarian cancer, malignant melanoma, high-grade brain tumors, and recurrent colorectal cancer (*3*). IQWIG also states that there is an urgent need for high-quality studies not only to assess the clinical benefit of PET but also to evaluate the diagnostic accuracy of PET. This assessment is expected to result in further restrictions on the use of PET/CT in Germany.

See Pages 64S and 77S

How did IQWIG come to these 2 conclusions that conflict with clinical practice in the United States and almost all other European countries? It is relatively easy to explain the statement that there is no evidence for a clinical benefit of PET and PET/CT. When PET was introduced into clinical oncology, a formal assessment of the clinical benefit of a diagnostic test was generally not required. More importantly, there is still no international agreement on how to define the clinical benefit of a diagnostic test. Randomized trials aiming to determine the impact of imaging on generally accepted hard clinical endpoints, such as overall survival, are prohibitively expensive. This is especially the case for an imaging probe such as $^{18}$F-FDG, which was developed by academia and is not patent-protected. As a consequence, lack of funding continues to prevent the large-scale randomized trials necessary to determine clinical benefit according to the criteria stipulated by IQWIG. In this supplement, Ware and Hicks (*2*) and Vach et al. (*4*) discuss in detail that these criteria may be inadequate and that more intelligent trial designs may allow us to assess the clinical benefit of PET at a fraction of the costs.

What is really surprising is that IQWIG also made the statement that there are insufficient data to determine the diagnostic accuracy of $^{18}$F-FDG PET in these common malignant tumors. In the last 20 years, the diagnostic accuracy of $^{18}$F-FDG PET has been evaluated in hundreds of clinical studies published in well-recognized scientific journals. How did IQWIG come to the conclusion that almost all these studies are of poor quality and do not allow evidence-based conclusions?

IQWIG uses the so-called Quality Assessment of Diagnostic Accuracy Studies (QUADAS) procedure to assess the quality of a publication on a diagnostic test (*3,5*). QUADAS asks 14 questions, which are listed in Table 1. All these questions address reasonable considerations when clinical studies on diagnostic tests are being planned. For example, the accuracy of a study evaluating $^{18}$F-FDG PET for differentiation of benign and malignant solitary pulmonary nodules is likely to be biased when the reader of the PET scans knows the results of histopathology (question 10). Conversely, the sensitivity of PET will be overestimated if histopathologic analysis is performed only in the case of a positive PET scan (question 6).

However, QUADAS is used very differently for the generation of IQWIG reports. For its reports, IQWIG typically gives grants to small companies specialized in preparing systematic reviews on various topics. For example, the company performing a review on PET in malignant melanoma for IQWIG has recently also performed a review on golimumab for ankylosing spondylitis, a review on manitol dry powder for inhalation for the treatment of cystic fibrosis, and an assessment to confirm an anaphylactic episode (*6*). These extremely diverse topics indicate that the reviewers do not judge the content of the reviewed publications but rather assess their quality solely by formal criteria as described by QUADAS. Each question raised by QUADAS is answered as yes, no, or unclear. The greater the number of questions answered with no or unclear, the lower is the quality of the publication. For such an approach to produce meaningful results, QUADAS would need to specify criteria that are a priori applicable to all types of diagnostic tests. Even a casual review of the criteria listed in Table 1 reveals that this is not the case for detection of distant metastases—the major application of $^{18}$F-FDG PET in oncology.

Received Oct. 31, 2011; revision accepted Nov. 1, 2011.

For correspondence or reprints contact: Wolfgang A. Weber, Department of Nuclear Medicine, University of Freiburg, Hugstetterstrasse 55, 79106 Freiburg, Germany.
E-mail: wolfgang.weber@uniklinik-freiburg.de

DOI: 10.2967/jnumed.111.100222

**TABLE 1**
Questions Asked by QUADAS Tool to Assess Quality of Studies Evaluating Diagnostic Tests

| Question no. | QUADAS tool question |
|---|---|
| 1 | Was the spectrum of patients representative of the patients who will receive the test in practice? |
| 2 | Were selection criteria clearly described? |
| 3 | Is the reference standard likely to correctly classify the target condition? |
| 4 | Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the 2 tests? |
| 5 | Did the whole sample or a random selection of the sample receive verification using a reference standard of diagnosis? |
| 6 | Did patients receive the same reference standard regardless of the index test result? |
| 7 | Was the reference standard independent of the index test (i.e., the index test did not form part of the reference standard)? |
| 8 | Was the execution of the index test described in sufficient detail to permit replication of the test? |
| 9 | Was the execution of the reference standard described in sufficient detail to permit its replication? |
| 10 | Were the index test results interpreted without knowledge of the results of the reference standard? |
| 11 | Were the reference standard results interpreted without knowledge of the results of the index test? |
| 12 | Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? |
| 13 | Were uninterpretable/intermediate test results reported? |
| 14 | Were withdrawals from the study explained? |

The reference standard to assess the presence or absence of metastatic disease is histopathologic analysis. If histopathologic analysis is not feasible (e.g., because a biopsy is associated with a high risk), follow-up imaging is generally accepted to prove or exclude metastases. Growth of a lesion on follow-up imaging is considered as evidence for metastatic disease, whereas lack of growth over a longer period excludes metastases. Both approaches to verify imaging findings are incompatible with QUADAS (see questions in Table 1). Histologic verification of a distant metastasis is not possible unless a lesion has been identified by an imaging study. As a consequence, the reference standard cannot be determined independently of the index test (question 6). Furthermore, the reference standard can be determined only when an imaging result is abnormal. Performing the reference standard in all patients would require an autopsy after the imaging study. Thus, the reference standard is not performed in the whole sample or a random selection of the sample (question 5). The pathologist analyzing a tissue sample is generally aware of this—that is, knows that there was an abnormality on an imaging study that resulted in the biopsy sample being analyzed. Consequently, the reference standard is not interpreted without knowledge of the results of the index test (question 11).

If follow-up imaging is used to prove or exclude metastatic disease, there are other conflicts with QUADAS. Follow-up needs to be done with imaging. This means that the index test becomes part of the reference standard (question 7). Furthermore, systemic therapies may affect the growth of metastases, and new metastases may develop during a longer follow-up period. This conflicts with questions 3 and 4. Overall, we are left with the conclusion that it is impossible to design a study for detection of distant metastases by an imaging modality without violating 6 of the 14 requirements made by QUADAS. Consequently, there can be no high-quality studies on the diagnostic accuracy of imaging for detection of distant metastases.

According to the widely accepted Fryback and Thornbury model for assessing the effectiveness of imaging studies (*7*), knowledge of the diagnostic accuracy of an imaging study is a prerequisite for assessing its diagnostic thinking efficacy, therapeutic efficacy, patient outcome efficacy, and social efficacy. Because QUADAS precludes high-quality studies on the diagnostic accuracy of an imaging test for detection of distant metastases, we are forced to conclude that such imaging tests cannot have a beneficial effect for the patient or for society. Consequently, all imaging for detection of distant metastases, be it ultrasound, CT, MRI, or PET, should be stopped immediately.

This reasoning according to the principles of evidence-based medicine is deliberately taken to the extreme to illustrate that highly unexpected conclusions can result from a purely formal analysis of scientific studies. There is abundant evidence that detection of distant metastases is highly beneficial to prevent complications (e.g., fractures due to bone metastases) and to avoid unnecessary surgery in patients with widespread metastatic disease. However, it is instructive to analyze why strict adherence to apparently reasonable principles described by 2 key papers (*5,7*) of evidence-based medicine leads to an absurd conclusion. QUADAS assesses whether a study is adequately designed to determine the sensitivity and specificity of a diagnostic test for a certain disease. Application of QUADAS to diagnostic tests for distant metastases shows that we cannot determine the sensitivity and specificity of such a test in an unbiased way. However, to determine whether an imaging test is efficient for detection of distant metastases, one does not need to know its exact sensitivity and specificity. It is sufficient to demonstrate that the test is more accurate than the current standard for testing for metastases. For example, if a patient is diagnosed with lung cancer we can start by comparing no staging examinations (sensitivity of 0% and specificity of 100% for detection of distant metastases) with physical examination. We can then compare physical examination with physical examination plus chest radiograph and so on. For all these comparisons, a reference standard is necessary only when the current standard and the new test yield different results. In these cases, an abnormality has been identified by 1 of the 2 modalities. Thus, discrepant findings can most of the time be clarified by histopathologic analysis. From these data, a comparison of the diagnostic accuracy of the 2 tests can be made, although we cannot determine the absolute sensitivity and specificity of the 2 tests. In other words, high-quality diagnostic studies can be made in complete violation of the principles of QUADAS.

The question is raised of how QUADAS was established and what the evidence is for its use in the formal assessment of the quality of diagnostic accuracy studies. The QUADAS tool was generated by a so-called Delphi procedure. The Delphi procedure (or method) describes a structured communication technique that uses questionnaires to achieve consensus among a panel of anonymous experts. The initial objective of the Delphi procedure was to predict developments in the field of science and technology. The name *Delphi* thus refers to the famous oracle of Delphi, who was consulted in ancient Greece before major endeavors such as the waging of wars or the founding of cities. An oracle and a consensus of experts are, however, strangely at odds with objective science and the principles of evidence-based medicine. In fact, expert opinion is generally considered as one of the lowest levels of evidence in evidence-based medicine and judged as insufficient for evidence-based decisions. So we have the paradox that in evidence-based medicine the level of evidence of diagnostic studies is assessed with a tool for which there is no sufficient evidence according to evidence-based medicine standards.

This simple example and the analysis by Ware and Hicks (*2*) show that the current use of evidence-based medicine and health technology assessments for imaging studies is untenable. At the root of the problem lies a lack of accepted trial designs to determine the diagnostic accuracy and the clinical benefit of imaging studies. A fundamental change of this situation requires that biostatisticians and "imagers" undergo a true Delphi procedure. The perhaps most famous inscription at the Apollo temple in Delphi was "ΓΝΩΘΙ ΣΑΥΤΟΝ – Know thyself" (*1*). Such a reflection on one's own deficiencies should make biostatisticians realize that the benefit of imaging studies can be more complex than expected and that seemingly reasonable generic models to assess the quality of imaging studies can lead to absurd conclusions. It is equally important that imagers reflect on their deficiencies in biostatistics and mathematic modeling. Only if the 2 disciplines realize and overcome their own limitations can there be the start of a fruitful collaboration to scientifically evaluate the benefits and limitations of modern imaging techniques.

**Wolfgang A. Weber**
*Department of Nuclear Medicine*
*University of Freiburg*
*Freiburg, Germany*

## REFERENCES

1. Plato. *Protagoras. 343ab.*
2. Ware RE, Hicks RJ. 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):64S–73S.
3. Projects page. IQWIG Web site. Available at: https://www.iqwig.de/language-selector.52.en.html?tid=1057&phlex_override_command=element. Accessed November 2, 2011.
4. Vach W, Høilund-Carlson PF, Gerke O, Weber WA. How to generate evidence for a clinical benefit of PET/CT in diagnosing cancer patients. *J Nucl Med.* 2011;52(suppl):77S–85S.
5. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol.* 2003;3:25.
6. Kleijnen Systematic Reviews Ltd. Web site. Available at: http://www.systematic-reviews.com. Accessed November 2, 2011.
7. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making.* 1991;11:88–94.