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?

Robert E. Ware and Rodney J. Hicks

Centre for Cancer Imaging, Peter MacCallum Cancer Centre, East Melbourne, Australia

Health technology assessment (HTA) has the objective of providing individual patients, clinicians, and funding bodies with the highest-quality information on the net patient benefits and cost effectiveness of medical interventions. Founded on systematic reviews of the available evidence, HTA aims to reduce bias and thereby provide a more valid evaluation of the benefits of new medical interventions than the primary studies themselves. Competing with the traditional role of medical experts, HTA agencies have gained considerable influence over public opinion and policy. The fundamental tenets of evidence-based medicine mandate that this influence should be used first and foremost for the benefit of patients. Over nearly 2 decades, multiple HTA systematic reviews in many countries have discredited most or all of the evidence pertaining to the ability of PET to improve patient-important outcomes. These determinations have delayed, restricted, and, in many cases, prevented access to this technology, especially by cancer patients. HTA systematic review findings are very much at variance with the opinion of clinicians. Our scrutiny of these reviews, benchmarking them against the core values of science and evidence-based medicine, has revealed errors of fact, inappropriate exclusion of pertinent data, and injudicious appraisal of the clinical relevance of evidence, potentially introducing bias into these reviews and compromising the validity of their conclusions about the net patient benefits of PET. We believe that our findings mandate that the molecular imaging community actively engage institutionalized HTA agencies to ensure appropriate representation of our primary data and adherence to the highest principles of evidence-based medicine.

Key Words: health technology assessment; PET; evidence-based medicine; cost effectiveness

J Nucl Med 2011; 52:64S–73S
DOI: 10.2967/jnumed.110.086611

We can think of no more important or immediate issue facing the nuclear medicine community than resolving the discordance between our profession's appraisal of our scientific integrity and judgments being made by institutionalized health technology assessment (HTA) agencies, especially with respect to the clinical value of PET. Of particular relevance to readers of *The Journal of Nuclear Medicine* is the routine claim of HTA analysts that there is little or no evidence to support the broadly held clinical judgment that PET and PET/CT improve outcomes of importance to patients. In the interest of our patients, we believe that this lack of consensus, which has endured for the best part of 2 decades, must be resolved.

The purpose of this article is to provide an evidence-based approach to assessing the validity of the HTA analysts' judgments. If most primary studies published in *The Journal of Nuclear Medicine* and other peer-reviewed scientific literature that promote the value of PET and PET/CT truly have little or no scientific merit, we must urgently remedy our research methodologies and decision-making processes.

BACKGROUND

A growing body of evidence produced by HTA organizations around the world alludes to insufficient evidence of the clinical utility of PET and PET/CT, affecting public perceptions of the ability of PET to improve patient-important outcomes. Many of these findings have been endorsed by officials of the International Network of Agencies of Health Technology Assessment, an umbrella organization for HTA groups operating in public health care (*1*). An example of the judgments that have characterized several of these HTA systematic reviews was published in October 2007 under the auspices of the United Kingdom's National Institute for Health Research (NIHR) Health Technology Assessment Program. A key finding of this review was that "there have been no clinical studies which demonstrate that FDG PET leads to an improvement in patient outcomes."

Health-care consumers and politicians in many countries have thus been led to believe that patients' and society's interests are being preserved by restraining access to PET

Received May 2, 2011; revision accepted Nov. 1, 2011.
For correspondence or reprints contact: Professor Rodney Hicks, Centre for Cancer Imaging, Peter MacCallum Cancer Centre, St. Andrew's Place, East Melbourne, VIC 3002, Australia.
E-mail: rod.hicks@petermac.org

pending development of sufficient scientific evidence of improved patient outcomes.

How Does the HTA View Align with the Judgments of Expert Clinicians About PET?

David Sackett, a leading proponent of evidence-based medicine (EBM), has stated that ". . .the overriding criterion for when to use a diagnostic test should be the usefulness of a given piece of diagnostic information to the clinician and to the patient. A useful diagnostic test does several things: it provides an accurate diagnosis, supports the application of a specific efficacious treatment, and ultimately leads to a better clinical outcome for the patient." His apt definition of what constitutes a useful diagnostic test has been embraced around the world by clinicians who have decided that PET is an invaluable tool for planning treatment of patients with many common and highly lethal malignancies. Unambiguous evidence of the benefits of PET in the routine care of individual patients has been supplemented by a growing body of largely consistent evidence published in peer-reviewed scientific literature (*2,3*). Increasingly, ^{18}F-FDG PET is also being used to monitor the efficacy of expensive and often-toxic therapies (*4*). This body of evidence informs clinicians of the direct impact PET can have, compared with alternative diagnostic approaches that have long been accepted to be beneficial and that are funded by health-care systems worldwide.

For example, in a previous supplement of *The Journal of Nuclear Medicine,* Shankar and Sullivan (*5*) of the National Cancer Institute's Cancer Imaging Program described PET as a "transformational technology." In 2010, the Clinical Oncologic Society of Australasia commented to the Australian Department of Health and Ageing, in response to a call for submissions on the process of evaluation of new technologies, that "PET has a critical and irreplaceable role in the investigation and management of several cancers,. . . supports informed decision making by patients and has advantages in terms of cost and patient quality of life by avoiding morbid interventions in patients for whom such treatment is not warranted."

In summary, there is generally strong clinical recognition of the utility of PET, particularly in oncology, which is at variance with many HTA systematic review conclusions.

How Has This Discord Developed?

In the past 20 years, there has been a seismic shift in the way clinical scientific knowledge is generated, legitimized, and disseminated. Clinicians were once trusted to determine the standards of excellent patient care. More recently, HTA analysts have become increasingly influential in this domain. Guided by principles developed within the framework of EBM, HTA groups use systematic reviews as an instrument to facilitate evidence-based health care (Table 1).

HTA systematic review, or secondary analysis of available evidence, is deemed to constitute a closer approximation to the truth than the individual primary studies, narrative reviews of expert clinicians, and editorial comment in peer-reviewed scientific journals. Viewed as the most valid information available, HTA systematic reviews have become increasingly important in guiding public policy, particularly with respect to funding new technologies, but also provide clinicians and

TABLE 1
Definitions of Terms Used

Term	Definition	Reference
Evidence-based medicine	"The conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research."	*36*
Evidence-based health	"A discipline centred on evidence based decision making about groups of patients, and populations, which may be manifest as evidence-based policy-making, purchasing or management."	*38*
Health-technology assessment	"A multidisciplinary field of policy analysis that studies the medical, social, ethical and economic implications of the development, diffusion and use of health technology."	*39*
Systematic review	"A scientific investigation in itself, with a preplanned Methods section and an assembly of original studies (predominantly randomised controlled trials and clinical controlled trials, but also sometimes, non randomised observational studies) as their subjects. The results of these multiple primary studies are synthesized by using strategies that limit bias and random error. These strategies include a comprehensive search of all potentially relevant studies and the use of explicit reproducible criteria in the selection of studies for review. Primary research designs and study characteristics are appraised, data are synthesized, and results interpreted."	*40*

patients with information that may be influential in guiding personal decisions on health-care options.

Although the aim of guiding clinical and funding decisions with the best available evidence is laudable, it is remarkable, especially given the outcome focus of EBM doctrine, that this profound change in the way the benefits of medical interventions to patients are judged has been adopted on the basis of logical induction rather than by validation with high-level evidence. If the process of secondary review does not reduce but rather increases bias, the potential to do more harm than good seems obvious, as does the risk that the ethical care of patients could be jeopardized (*6*).

What Are the Important Aspects of Our Critical Appraisal of PET HTA Systematic Reviews?

To assess whether the systematic review process has produced judgments that are closer to the truth than the primary studies and other data that are available for appraisal, we have scrutinized influential HTA systematic reviews published between 1996 and 2010 (Table 2). In particular, we sought evidence that the secondary review process has introduced bias about the merits of existing evidence.

The Standards for Reporting Studies of Diagnostic Accuracy (STARD) (*7*), the Quality Assessment of Diagnostic Accuracy Studies (*8*) initiatives, and the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, which is, as yet, incomplete (http://srdta.cochrane.org/), are excellent EBM publications to guide judgments about the validity of quality filters applied to primary studies. They also provide a temporal foundation for judging the quality of HTA systematic review decisions, because reviews before 2003 relied on assessment constructs developed for systematic reviews of therapeutic interventions and were poorly validated in the context of diagnostic tests.

However, these guides do not provide an adequate framework for our appraisal. They relate only to diagnostic accuracy assessments, whereas the greatest divergence between clinical experts and HTA systematic review analysts about the value of PET generally involves judgments about the relationship between the acknowledged incremental accuracy of PET and improved patient outcomes. These guides are also process-based and not outcome-based. Several instruments designed to guide critical appraisal of systematic reviews of evidence (*9*) (SIGN-www.sign.ac.uk/guidelines/fulltext/50/checklist1.html) have similar limitations.

Our scrutiny of HTA systematic reviews has been conducted within the framework of an adaptation of Australia's National Health and Medical Research Council's 2009 guidelines for assessing a body of evidence and formulating guidelines (www.nhmrc.gov.au__files_nhmrc_file_guidelines_evidence_statement_form.pdf). Our schema for assessing the various HTA systematic reviews is shown in Table 3. We have not taken a balance sheet approach to judging the worth of HTA systematic reviews by weighing good against bad in some subjective and arbitrary decision-making process. Rather we have adopted a cutoff approach because, within our logical construct, HTA systematic reviews that do not undertake judicious, explicit, and conscientious evaluations of the existing evidence have a low probability of providing a closer estimate of the truth than the primary evidence itself.

Although our interest in the HTA process was stimulated by our personal experience with PET reviews performed in Australia, which we have previously documented (*10,11*), we believe that issues raised are highly relevant to the global medical community, including that of the United States. Although broadly based funding for PET has been achieved, particularly in response to data from the National Oncologic PET Registry process (*12*), negative international evidence-based reviews of PET may, in the future, increase pressure to reverse these decisions if health-care expenditure continues to rise and requires budgetary constraint.

Did HTA Systematic Reviews Published Between 1996 and 2010 Make Judicious Appraisals of the Strength of Evidence Pertaining to PET's Ability to Improve Patient-Important Outcomes?

Detailed review of the published HTA systematic reviews has identified material flaws in each, as summarized in

TABLE 2
HTA Systematic Reviews

Health Technology Assessment Agency	Year of review publication and country of origin	Abbreviation
Veterans Affairs Technology Assessment Program	1996, United States	VATAP
Medicare Service Advisory Committee	2001, Australia	MSAC 2001
Danish Centre for Evaluation and Health Technology Assessment	2001, Denmark	DACEHTA
Institute for Clinical Evaluative Sciences	2001, Canada	ICES
Health Technology Board for Scotland	2002	HTBS
National Institute for Health Research	2007, United Kingdom	NIHR
Medical Services Advisory Committee (PET in Recurrent Colorectal Cancer)	2008, Australia	MSAC CRC 2008
Medical Services Advisory Committee (PET for Head and Neck Cancer)	2009, Australia	MSAC H&N 2009
Medical Services Advisory Committee (PET for Lymphoma)	2010, Australia	MSAC Lymphoma 2010

TABLE 3
Framework for This Critical Appraisal of Health Technology Assessment Systematic Reviews

Question	Subquestion
Were factual data represented faithfully?	
Is there evidence of judicious appraisal of the "strength" of the evidence base?	Has there been injudicious appraisal of the quality and quantity of available evidence?
	Is there evidence of injudicious appraisal of the "consistency" of the evidence base?
	Is there evidence of injudicious appraisal of the clinical importance of the evidence?
Is there evidence that individual patient outcomes were paramount?	

Table 4. Documentary evidence of these errors has been assembled through a combination of freedom-of-information requests, direct contact with members of the review panels, and independent analysis of the literature that was both reviewed and discarded in the process of HTA systematic review. Many of our findings have been supplied to the primary authors of these reviews for rebuttal or correction without success. For the sake of brevity, we provide only a small fraction of the examples that we have assembled of failures to meet the requirements of meticulous scientific review.

Examples of Erroneous Representation of Factual Data. Errors of fact constitute the most potent source of bias in scientific reports and must be corrected if identified. We have found evidence of material factual errors in several influential HTA systematic reviews scrutinized in this evaluation and include misrepresentation both of primary data and of the authors' conclusions.

Some of the factual errors in Medical Services Advisory Committee (MSAC) 2001 have been published previously (*10*). In its evaluation of the role of ^{18}F-FDG PET in the staging of lung cancer, the Danish Centre for Evaluation and Health Technology Assessment (DACEHTA) stated incorrectly that the study of Pieterman et al. (*13*) was a randomized controlled trial and that it examined cost efficacy, and DACEHTA incorrectly asserted that Marom's study (*14*) was unblinded when blinding was clearly described in the study methodology. DACEHTA incorrectly stated that the study of Saunders et al. (*15*) included 17 malignant patients and 67 benign patients when the study in fact included "84 patients with biopsy proven lung cancer and 13 with a high suspicion of lung cancer." DACEHTA's errors are admittedly identified in an unofficial translation from Danish. However, the same translation formed a primary source of evidence for the Health Technology Board for Scotland (HTBS) evaluation of PET, which included an extensive verbatim presentation of the DACEHTA findings. DACEHTA's findings have also been referenced by several other HTA systematic reviews.

Further, NIHR erroneously categorized the PET study of Porceddu et al. (*16*) in restaging patients with SCC of the head and neck as one of diagnostic accuracy. The clearly stated aim of this study was to evaluate the incremental management impact and therapeutic decision-making

TABLE 4
Evidence of Flaws Within Reviewed Health Technology Assessment Systematic Reviews

Review	Factual errors	Injudicious appraisal of quality and quantity of evidence	Injudicious appraisal of evidence consistency	Injudicious appraisal of clinical importance of evidence	Individual patient outcomes not paramount
Pre-STARD					
VATAP	+*	+*	+	+*	+*
MSAC 2001	+	+	+	+	+*
ICES		+*	+	+	
DACEHTA	+*	+	+	+	+*
HTBS	+	+	+*	+	+*
Post-STARD (2003)					
NIHR	+*	+	+	+	+*
MSAC CRC 2008	+	+*	+	+	
MSAC H&N 2009	+*	+	+	+	
MSAC Lymphoma 2010		+*	+	+	

*Examples detailed in text.

potential of PET in a specific clinical context. Because of this aim, spectrum bias precluded derivation of the usual diagnostic test metrics such as sensitivity and specificity. Nevertheless, NIHR reinterpreted the available published data erroneously, stating that "PET's sensitivity was five out of ten (50%)." This misrepresentation of the facts would create the impression that PET is not an accurate or useful test, especially as NIHR does not reference Porceddu's conclusion, as follows: "Patients who have achieved a complete response at the primary site but have a residual anatomic abnormality in the neck that is negative on PET scan approximately 12 weeks after treatment do not require neck dissection and can be safely observed."

MSAC Head and Neck Cancer 2009 used NIHR as a starting point for its own analysis. MSAC replicates the NIHR errors, adding several inaccuracies of its own. The study of Porceddu et al. is classified as retrospective when prospective data collection was clearly documented. Claims that no data were provided on the accuracy of PET for detecting distant metastases in this population conflict with the detailed verification procedure documented in the original publication. These inaccuracies were material to MSAC's classification of this primary study as being of only fair quality.

Examples of Injudicious Appraisal of the Quality or Quantity of Available Studies. Before the publication of STARD guidelines, HTA systematic reviews were potentially constrained by poorly validated quality filters for assessing the risk of bias in primary studies of PET. Yet before publication of these criteria, the quality filters adopted by many PET HTA systematic reviews were not applied in a judicious, explicit, and conscientious manner. Nevertheless, these HTA systematic reviews are widely referenced and were often incorporated into later reviews.

The quality filters developed for the Veterans Affairs Technology Assessment Program (VATAP) review were subsequently influential in several later HTA systematic reviews (e.g., DACEHTA and Institute for Clinical Evaluative Sciences [ICES]). To illustrate failures inherent in the application of VATAP's grading system, we cite the 1995 paper of Valk et al. (*17*) examining the value of PET in presurgical staging of patients with non–small cell lung cancer (NSCLC). This study was graded as D (on an A–D scale for quality), suggesting multiple serious methodologic flaws with a high risk of bias. However, the summary table records a + against each of 3 quality filter items, signifying compliance. VATAP evaluators claimed the final diagnosis was not determined independently of the PET scan result, yet the authors indicate that histology was the reference standard for all 76 patients in whom the accuracy of PET and CT for detecting mediastinal nodal metastases was assessed. This validation was also supplemented by clinical and imaging follow-up. VATAP claimed that comprehensiveness of the nodal sampling was unclear, when in fact the paper contains very detailed descriptions. VATAP indicated that failure to enroll an equal number of patients with and without cancer indicated methodologic weakness. One can assess the merits of VATAP's decision by reference to a contemporary publication (*18*) that advises, "Almost any test can distinguish the healthy from the severely affected; this ability tells us nothing about the clinical utility of a test. The true, pragmatic value of a test is therefore established only in a study that closely resembles clinical practice."

In a compounding factual error, VATAP incorrectly stated that the study included 76 patients rather than the true number, 99. The discussion of Valk et al. emphasized the clinical importance of the finding that PET uniquely detected distant metastases in 11 patients, yet VATAP described the data of Valk et al. as anecdotal and excluded the information from evidence summary tables because of small patient numbers. It seems probable that multifactorial misjudgment of Valk's primary study resulted in this seminal high-quality data, detailing the potential patient benefits of whole-body PET in NSCLC, being unfairly judged as having little or no scientific value by VATAP authors.

ICES adopted a modification of the VATAP 4-point scale for grading the quality of primary studies. An "a priori decision to concentrate on A and B grade articles" was made, presumably to reduce risk of bias. Otherwise relevant primary studies of PET in oncology, not judged to be A or B grade, were not referenced in any way in the report, creating the impression that these studies did not exist. ICES (Susan Garfinkel, written communication, 2005) provided us with the reviewers' summary score sheets after we questioned the validity of their decisions in relation to published primary studies from our institution. Of the multiple peer-reviewed articles from our group, ICES graded only 12. Of these, 11 were subsequently excluded, including all 3 published in *The Journal of Nuclear Medicine* (*19–21*). Despite the ICES protocol requiring several or multiple flaws to attract C or D scores, respectively, only a single flaw was ascribed to 9 of the 11 excluded articles. This fault was a purported lack of independence between PET and the reference standard. However, all the excluded studies used a compound reference standard including pathology, if available, correlative imaging, and a long period of follow-up. Since none of this information was available at the time of PET scan reporting, the PET scans were, by definition, interpreted without knowledge of the reference standard and therefore could not have been subject to bias. Another primary study in which lack of independence was deemed to constitute a high risk of bias (*22*) was a cohort study that compared survival in 2 groups of NSCLC patients, one whose chemoradiation therapy was planned with and the other without PET. When PET was used to plan radiation treatment, that process could not have taken place with knowledge of the reference standard, which was the duration of the patient's subsequent survival.

In contrast, an abstract pertaining to a randomized trial of PET in NSCLC was awarded an A for quality. This evidence was included in the evidence tables and discussion

even though the limited detail in the abstract precluded thorough assessment of methodologic quality. The final published results of this trial (*23*) will be discussed in more detail below.

The development of new standards for review of the diagnostic imaging literature has not prevented subsequently published HTA systematic reviews from including unreasonable assessments of the strength of the available evidence base. MSAC CRC 2008 failed to evaluate the paper of Kalff et al. examining the clinical impact of PET in restaging colorectal cancer (*24*) even though the primary study was cited. The methodology that underpinned this study subsequently formed the basis for a scientific study protocol that MSAC endorsed in 2003 as part of its National PET Data Collection process, with the paper arising from that multicenter study being subsequently published in *The Journal of Nuclear Medicine* (*25*). This study was discussed in great detail within this HTA systematic review, whereas the earlier paper was ignored.

When questioned about this omission, MSAC claimed that it had missed this reference because the reference was not cited in the "high quality" NIHR 2007 review that was used as a starting point to identify prior publications. However, Kalff's primary study was described in NIHR's discussion section as a high-quality prospective study of a consecutive series of patients that was well designed and well documented. When MSAC was made aware of the erroneous omission, it indicated that the exclusion of the primary study was not pertinent to its conclusions. MSAC CRC 2008 also omitted a cost consequence analysis, commissioned by a grant from the Department of Health and Aging, which demonstrated that PET could reduce health-care costs by approximately $4,000 per patient by virtue of avoidance of inappropriate surgery. Although this analysis was an internal document, this omission is inexplicable given that the study's chief author was also an author of MSAC CRC 2008 and that EBM places great value on finding and reviewing unpublished studies (*26*).

MSAC CRC 2008 also identified a systematic review of PET in recurrent colorectal cancer, authored by Dietlein et al. (*27*), that included 15 primary studies and hundreds of patients. The accuracy of PET was one of the parameters appraised, and pooled values for sensitivity, specificity, and likelihood ratios for PET and CT (as the clinical comparator) were detailed. PET was found to have superior sensitivity and specificity, with minimal overlap of the 95% confidence levels and markedly superior positive and negative likelihood ratios. This high-level evidence was omitted from the "Accuracy" discussion in MSAC's executive summary, which details positive predictive values from only 4 primary studies. Similarly, the executive summary does not refer to the systematic review finding of Dietlein et al. that management change occurred in 34% of patients (95% confidence interval, 31%–38%), mentioning only a single primary study and an unpublished randomized controlled trial. MSAC's justification for this omission of apparently relevant, high-quality evidence from its conclusions is that Dietlein et al. evaluated comparative accuracy, whereas MSAC indicated that they were addressing the question of incremental accuracy. Ironically, this is an approach that MSAC had specifically criticized in studies from our institution and was given by ICES reviewers as a reason for excluding our publications from analysis in their review.

Improved reporting is evident in MSAC Lymphoma 2010, with the adoption of the guidelines of the Quality of Reporting of Meta-Analyses conference (*28*). These allowed some quantitative understanding of the potential evidence base. There were also rudimentary qualitative reasons given for evidence exclusion. Incorporation of results from the MSAC 2001 HTA systematic review led to the inclusion of 38 publications before 2000. An updated literature search was then performed, extending from 2000 to 2009. This identified 2,319 potentially relevant studies. Of these, 2,107 (91%) were excluded for a variety of reasons without publication retrieval. Of 212 remaining studies selected for detailed analysis, only 16 (0.7% of the original sample), plus an unpublished report from the Australian Data Collection Study, were included in the HTA systematic review. Five systematic reviews and an HTA (NIHR) were also mentioned (but not HTBS or ICES). Although the MSAC 2001 systematic review was graded as a high-quality systematic review, the findings pertaining to the 38 studies reviewed were not detailed in the executive summary or results section. Furthermore, neither this evidence nor other systematic reviews or HTA appraisals were added to the evidence summary matrices.

That such a large number of primary studies and other evidence were excluded in this review raises questions about the validity of the analysis to clinical practice. Importantly, this failing was specifically remarked on in the final report under the heading "Expert Opinion" as follows: "Many or most of the excluded studies indicate a positive impact of PET on staging, response assessment or prognostic evaluation, so that the clinical utility of PET may have been underestimated.. . .Several studies which failed to meet the inclusion criteria have shown a dramatic relationship between early PET response and outcome."

As an example, the multicenter study of Gallamini et al. detailed 260 patients with Hodgkin lymphoma, demonstrating a highly significant relationship between treatment outcome and ^{18}F-FDG PET scan results after 2 cycles of chemotherapy (*29*). Multivariate analysis showed PET status to be the only significant predictor of treatment outcome ($P < 0.0001$). PET was found to have a 93% positive predictive value and 92% negative predictive value for 2-y progression-free survival. However, MSAC excluded this study on the basis of the comparator used for assessment. Presumably, this was because the primary study used the International Prognostic Score and not simply CT as the comparator for PET. However, the exact reason is not stated explicitly enough for certainty. Similarly, the 2006 study by

Zinzani et al. (*30*) examining the same clinical question, with concordant findings, was also excluded by MSAC for the same reason.

Similar errors of omission are apparent across most of the HTA systematic reviews that we have evaluated. That such a large body of the published literature is routinely excluded from evaluation in HTA systematic reviews is a matter of concern regarding either the experimental methods accepted by major peer-reviewed journals such as *The Journal of Nuclear Medicine* or, alternatively, the process by which evidence is evaluated within the EBM process by HTA agencies.

Evidence of Injudicious Appraisal of the Consistency of the Evidence Base. The primary studies of PET are remarkable for their consistency in demonstrating superior diagnostic accuracy in relation to contemporary clinical comparators (*2*). These data have been recapitulated and, in many cases, improved by the introduction of PET/CT (*3*). In many of the papers cited, clinicians have also documented or imputed marked management changes.

This consistency receives little emphasis in the HTA systematic review appraised. Rather, undue emphasis is placed on the risk of bias, thereby questioning the validity of the primary study. For example, with respect to the value of PET for staging of NSCLC, HTBS notes that 33 papers were identified, with the commentary that "almost all reported that PET is more accurate and more sensitive than CT," and 2 positive meta-analyses were also cited. These statements were, however, downplayed by the comment, common to many other HTA systematic reviews, that ". . .these results need to be confirmed in larger randomised trials."

The HTA systematic reviews also stress apparent inconsistencies between studies. For example, although HTBS noted a published randomized controlled trial (*31*) reporting a statistically significant ($P = 0.003$) reduction in patients undergoing futile thoracotomies in the PET group, this was discounted under the heading "Limitations of the Evidence" by reference to an apparently "conflicting" randomized controlled trial available only in abstract form at that time. Assessment of quality of this trial was not possible, as not all patients had been fully evaluated. Indeed, detailed review of the subsequently published primary data from this trial (*23*) indicates spectrum bias, with most patients having very early stage disease. Also, the primary endpoint was the number of thoracotomies avoided in patients randomized to PET staging, and yet thoracotomy was frequently performed even in patients for whom PET made accurate incremental findings of mediastinal nodal and distant metastatic disease. If the institution's standard of care for NSCLC was thoracotomy irrespective of disease stage, the ethical and scientific context for performing the randomized controlled trial could be questioned.

A major benefit claimed for systematic reviews is to resolve uncertainty when original research, reviews, and editorials disagree (*32*), yet PET and PET/CT studies are remarkably consistent in supporting the clinical utility of these techniques. Paradoxically, HTA systematic reviews commonly follow the HTBS example of emphasizing apparent heterogeneity as a justification for calling for more definitive evidence. However, when evaluated in a clinical context, variability in results reflects appropriate clinical selection of a patient population to address a specific clinical question.

Evidence of Injudicious Appraisal of the Clinical Importance of Evidence. Injudicious appraisal of the strength and consistency of the evidence base that we have pointed to must introduce a risk that HTA systematic review judgments about the clinical importance of the evidence will also be biased. One threshold question that has not been judiciously addressed when HTA systematic reviews were formulating their decisions about the merits of PET is as follows: accepting that estimates of the diagnostic performance of PET contain some bias, what is the likelihood that the consistently higher accuracy of PET and the 30%–40% documented or imputed change in management related to more accurate disease classification is really a null finding?

Injudicious assessment of the clinical impact of the primary study arises because several HTA systematic reviews, both before STARD (VATAP and ICES) and after STARD (NIHR), fail to ask a clinically focused question specifically addressing patient-important outcomes. For example ICES simply states, "We examined the evidence for clinical applications of PET among 7 commonly occurring categories of cancer." NIHR 2007 replicates this failure in the post-STARD era, stating its research questions as, "What is the clinical effectiveness of FDG PET for the management of the following cancers. . .?". NIHR's conclusion that, as of 2007, there were no published studies demonstrating that ^{18}F-FDG PET leads to improved patient outcomes is potentially a biased appraisal of the evidence that, in part, relates to the impossibility of providing a valid judgment of the clinical relevance of the evidence across such a wide spectrum of cancers and clinical management questions.

We believe that injudicious appraisal of the evidence pertaining to the clinical importance of PET in the context of individual patients is intimately related to the differences between HTA systematic review methodologists and clinicians regarding the types of evidence required to make valid judgments about the patient benefits of diagnostic tests. In short, HTA analysts have remained highly skeptical of the clinical view that studies validating diagnostic accuracy with a low risk of bias can be used to make sound judgments about potential patient benefit, if reduction in false-negative and false-positive results can be imputed (or shown in trial-based evidence) to have a direct and material impact on patient-important outcomes. Clinicians intuitively understand that false-negative results can delay appropriate introduction of effective treatments or result in patients receiving inappropriate treatment based on inadequate knowledge of the presence or extent of disease,

whereas false-positive results often lead to unnecessary anxiety, institution of sometimes morbid and expensive investigations and therapy, or denial of potentially curative treatment.

Convergence to the clinical view that randomized controlled trials may be neither ethical nor essential for judging the patient benefits of diagnostic technologies appears to be gradually occurring among EBM and HTA methodologists (*33*). The Grading of Recommendations Assessment, Development and Evaluation Working Group (*34*) still considers randomized studies to be the ideal method for comparing diagnostic approaches but recognizes that observational studies comparing alternative diagnostic strategies with assessment of direct patient-important outcomes can be the basis for strong recommendations about the patient benefits of a more accurate diagnostic test.

Although the continuing evolution of the EBM literature toward a greater focus on patient-important outcomes is encouraging, most HTA systematic reviews continue to negatively judge the evidence supporting PET for want of evidence from randomized controlled trials. As an example, MSAC CRC 2008, referring to the potential for improving patient outcomes related to the demonstrated ability of PET to produce a significant number of additional true-positive findings for extrahepatic metastases in comparison to best conventional evaluations, comments, "In the absence of randomised controlled clinical trials of alternative treatments in this patient group, it is not known whether the quality of life benefits from avoiding surgery and instigating alternative management outweigh any potential benefit of surgery in providing local disease control."

Yet, in the context of contemporary standards of clinical care of hepatic recurrence of colorectal cancer, extrahepatic metastases (detected clinically or by conventional imaging) usually preclude attempted curative surgery because there is no reliable evidence of improved outcomes in this situation. MSAC acknowledges PET to be safe, and also quantifies the harm associated with hepatic resection as a death rate of 1%–2% and a catastrophic morbidity of approximately 10%. With the information available, it seems an inescapable conclusion that attempted hepatic resection will do the patient more harm than good if performed on patients for whom PET has uniquely defined extrahepatic metastases, yet MSAC creates unnecessary uncertainty by emphasizing the lack of data from randomized controlled trials. As a consequence of this approach, the evidence supporting PET's beneficial effect in this context is said to be supported only by expert opinion.

The NIHR judgment that there are no published studies demonstrating that ^{18}F-FDG PET improves patient outcomes rests heavily on discrediting the clinical relevance of the many nonrandomized controlled trial–based primary studies and systematic reviews that were evaluated, as well as discounting completely the randomized controlled trial of van Tinteren et al. (*31*) that clearly shows improved patient outcomes.

Evidence That Factors Unrelated to Patient-Important Outcomes Risked Introducing Bias into HTA Systematic Reviews. EBM values mandate that evidence evaluation be conducted first and foremost from the perspective of individual patients. However, several HTA systematic reviews (VATAP, MSAC 2001, HTBS, and NIHR) adopted appraisal protocols that give societal benefit greater value than individual-patient benefit. For example, both MSAC 2001 and NIHR used the hierarchy of diagnostic efficacy established by Fryback and Thornbury (*35*) to quantify the level of evidence available. This hierarchy is a 6-point scale that accords the highest value to demonstration of societal benefit as defined by cost efficacy, cost benefit, or cost utility. Studies designed to assess the diagnostic accuracy and predictive value of PET were relegated to levels 2 and 3 under that schema. HTA systematic reviews of PET repeatedly judged such studies as low-level evidence, even though they contain knowledge that is crucial to decision making in the context of individual patient management and therefore of paramount importance within a true EBM context. We believe this decision about classifying the value of evidence has significant potential to prejudice judgments about the strength consistency and clinical relevance of evidence in relation to patient-important outcomes.

Contrary to the tenets of EBM that require the needs of individual patients to be paramount (*36*), it appears that at least one of the reviews was structured to produce findings and recommendations that were beneficial to those with responsibility for funding PET. Although MSAC was established to be an "independent multidisciplinary scientific committee," documents obtained under Freedom of Information Legislation show that for MSAC 2001 "the objective was to retain funding at the current level."

DISCUSSION

One of the many roles of peer-reviewed literature is to guide clinicians and patients on the merits of medical interventions. *The Journal of Nuclear Medicine* has an ethical responsibility to its readers and their patients to provide the best possible information. The editors and reviewers of articles submitted for publication are charged with responsibility for ensuring that the highest possible scientific standards are met. Key assurances are required from the authors that the work was performed within an appropriate ethical framework, that they stand by the primary data presented, and that they agree with the conclusions reached. They are also asked to identify potential conflicts of interest. The paper is then subjected to scrutiny regarding the importance and relevance of the research question, the appropriateness of the literature review that provides a background to the study, the rigor of the methodology (including recognition and exclusion, where possible, of potential sources of bias), the acquisition of sufficient and appropriate experimental data to address the problem at hand, the application of appropriate analysis approaches to the data acquired, and a balanced discussion of the findings. Most

importantly, the conclusions reached need to be supported by the data presented.

Despite the best efforts of authors, editors, and reviewers, papers with significant scientific flaws unfortunately do get published. Such errors are, however, subject to further scrutiny within the scientific community. The broader readership is given an immediate right of reply to point out errors of fact, omission, or commission in correspondence with the editor and, through the editorial office, with the study authors. The authors then have an opportunity to defend or correct any perceived errors. When significant faults are identified, errata are published or the paper can be withdrawn. The second level of audit is the recapitulation of similar confirmatory studies by other groups. Through this incremental process, new scientific truths are approximated ever more closely. A key complement to the publication of primary data is the collation of the broader published results, supplemented by expert opinion based on individual experience, to give a synthesis of the available information.

It is this latter process—wherein personal opinion and selection of only those papers that support that opinion have the potential to bias the conclusions presented—that has been the target of institutionalized EBM. However, if the laudable desire to present unbiased opinions is not accompanied by the same scientific rigor and audit that underpin the publication of the primary studies, one opinion biased from a perspective of clinical experience is simply transposed with another opinion biased by other perspectives, most prominently the desire to restrict health-care expenditure for the benefit of "society."

Our own experience, supported by documentary evidence, with HTA systematic reviews has convinced us that there is a fundamental asynchrony between the key quality checks applied within the peer-reviewed literature and those practiced within many of these reviews (Table 5). Institutionalized EBM has sought to occupy the scientific high ground and regularly dismisses the value and validity of a large component of our collective efforts, considering the papers published in high-impact journals such as *The Journal of Nuclear Medicine* to be of doubtful scientific merit because of bias. However, we believe that we have provided credible evidence (and a great deal more such evidence exists) that the systematic reviews of several international HTA groups have introduced significant bias within their own secondary appraisals of the evidence relevant to the patient benefits of PET in oncology. The findings and recommendations of these reviews therefore cannot represent a closer approximation to the truth than do the primary studies, nor do they invalidate the summations of evidence by members of our peer group that have been published in our journals. By virtue of the values and belief systems inherent in EBM, encouraging patients (and health-care administrators) to adopt recommendations about PET that are prejudiced is likely to cause more harm than good.

Our belief is that the advice of clinical experts has been greatly undervalued in the overall appraisal of the effect of PET on patient-important outcomes, both from the perspective of bottom-up clinical judgment relating to individual patient's experience and from the perspective of the many judgments required within the top-down analyses of primary evidence in systematic reviews. We believe that the molecular imaging community has a responsibility to challenge the misinformation arising from prejudiced assessments of the evidence base relating to PET and claims that patients and society have benefitted by restricting access to this technology. However, we also acknowledge that HTA has an important role in the context of constrained health resources, and we applaud the motivation shown to

TABLE 5
Comparison of Peer-Reviewed Literature and HTA Systematic Review Processes

Process	Peer-reviewed literature	HTA systematic review
Acknowledgment of authorship, validity of data and conclusions	Required to be signed by all authors before review	Evidence that dissent among clinical experts on panels was often not captured or acknowledged
Conflict-of-interest statement	Required and transparent	Literature reviewers often contracted by governments or third-party insurers, leading to potential conflict of interest
Cogent and appropriate articulation of research question	Pivotal	Often poorly defined or not explicitly stated
Literature review	All relevant prior art should be cited and is assessed by reviewers and readership	Filters placed on what was evaluated
Data acquisition	Explicit methodology that allows reproduction is mandated	Reasoning seldom provided for including or excluding primary data
Analysis	Established statistical methods	Varied but largely qualitative
Conclusions	Patient-focused	Society-focused
Responsiveness to criticism	Enshrined in process	Resisted

improve the methodologies used in assessing new technologies. Stronger engagement between clinician groups and an iterative process that emulates many of the beneficial features of the peer-review process are encouraged.

Hopefully, this article will stimulate a robust discussion among protagonists that will lead to genuine improvements in the quality of both the primary scientific literature and the secondary reviews thereof. Cancer patients deserve no less.

ACKNOWLEDGMENT

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Adams EJ, Almazan C, Morland B, Bradbury I, King R, Rheinberger P. Joint project of the International Network of Agencies for Health Technology Assessment: part 2—managing the diffusion of positron emission tomography with health technology assessment. *Int J Technol Assess Health Care.* 2006;22:149–154.
2. Gambhir SS, Czernin J, Schwimmer J, Silverman DH, Coleman RE, Phelps ME. A tabulated summary of the FDG PET literature. *J Nucl Med.* 2001;42 (suppl):1S–93S.
3. Czernin J, Allen-Auerbach M, Schelbert HR. Improvements in cancer staging with PET/CT: literature-based evidence as of September 2006. *J Nucl Med.* 2007;48(suppl 1):78S–88S.
4. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med.* 2009;50(suppl 1):122S–150S.
5. Shankar LK, Sullivan DC. PET/CT in cancer patient management. *J Nucl Med.* 2007;48(suppl 1):1S.
6. Leeder SR, Rychetnik L. Ethics and evidence-based medicine. *Med J Aust.* 2001;175:161–164.
7. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med.* 2003;138:W1–W12.
8. 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.
9. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. *JAMA.* 1994;272:1367–1371.
10. Ware RE, Francis HW, Read KE. The Australian government's review of positron emission tomography: evidence-based policy-making in action. *Med J Aust.* 2004;180:627–632.
11. Hicks RJ. Editorial: health technology assessment and cancer imaging—who should be setting the agenda? *Cancer Imaging.* 2004;4:58.
12. Coleman RE, Hillner BE, Shields AF, et al. PET and PET/CT reports: observations from the National Oncologic PET Registry. *J Nucl Med.* 2010;51:158–163.
13. Pieterman RM, van Putten JW, Meuzelaar JJ, et al. Preoperative staging of non-small-cell lung cancer with positron-emission tomography. *N Engl J Med.* 2000;343:254–261.
14. Marom EM, McAdams HP, Erasmus JJ, et al. Staging non-small cell lung cancer with whole-body PET. *Radiology.* 1999;212:803–809.
15. Saunders CA, Dussek JE, O'Doherty MJ, Maisey MN. Evaluation of fluorine-18-fluorodeoxyglucose whole body positron emission tomography imaging in the staging of lung cancer. *Ann Thorac Surg.* 1999;67:790–797.
16. Porceddu SV, Jarmolowski E, Hicks RJ, et al. Utility of positron emission tomography for the detection of disease in residual neck nodes after (chemo) radiotherapy in head and neck cancer. *Head Neck.* 2005;27:175–181.
17. Valk PE, Pounds TR, Hopkins DM, et al. Staging non-small cell lung cancer by whole-body positron emission tomographic imaging. *Ann Thorac Surg.* Dec 1995;60:1573–1581.
18. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA.* 1994;271:389–391.
19. Hicks RJ, Kalff V, MacManus MP, et al. ^{18}F-FDG PET provides high-impact and powerful prognostic stratification in staging newly diagnosed non-small cell lung cancer. *J Nucl Med.* 2001;42:1596–1604.
20. Hicks RJ, Kalff V, MacManus MP, et al. The utility of ^{18}F-FDG PET for suspected recurrent non-small cell lung cancer after potentially curative therapy: impact on management and prognostic stratification. *J Nucl Med.* 2001;42:1605–1613.
21. Kalff V, Hicks RJ, Ware RE, Hogg A, Binns D, McKenzie AF. The clinical impact of ^{18}F-FDG PET in patients with suspected or confirmed recurrence of colorectal cancer: a prospective study. *J Nucl Med.* 2002;43:492–499.
22. Mac Manus MP, Wong K, Hicks RJ, Matthews JP, Wirth A, Ball DL. Early mortality after radical radiotherapy for non-small-cell lung cancer: comparison of PET-staged and conventionally staged cohorts treated at a large tertiary referral center. *Int J Radiat Oncol Biol Phys.* 2002;52:351–361.
23. Viney RC, Boyer MJ, King MT, et al. Randomized controlled trial of the role of positron emission tomography in the management of stage I and II non-small-cell lung cancer. *J Clin Oncol.* 2004;22:2357–2362.
24. Kalff V, Hicks R, Ware R, Binns D, McKenzie A. F-18 FDG PET for suspected or confirmed regional recurrence of colon cancer: a prospective study of impact and outcome. *Clin Positron Imaging.* 2000;3:183.
25. Scott AM, Gunawardana DH, Kelley B, et al. PET changes management and improves prognostic stratification in patients with recurrent colorectal cancer: results of a multicenter prospective study. *J Nucl Med.* 2008;49:1451–1457.
26. Jackson T, Hicks RJ, MacManus MP. *A Cost Consequence Study of Care Following Cancer Staging With and Without the Use of F-18 FDG PET Scanning: Final Report to the Consultative Committee on Diagnostic Imaging*. Woden, Australia: Department of Health and Ageing; 2003.
27. Dietlein M, Weber W, Schwaiger M, Schicha H. ^{18}F-Fluorodeoxyglucose positron emission tomography in restaging of colorectal cancer [in German]. *Nuklearmedizin.* 2003;42:145–156.
28. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-Analyses. *Lancet.* 1999;354:1896–1900.
29. Gallamini A, Rigacci L, Merli F, et al. The predictive value of positron emission tomography scanning performed after two courses of standard therapy on treatment outcome in advanced stage Hodgkin's disease. *Haematologica.* 2006;91:475–481.
30. Zinzani PL, Tani M, Fanti S, et al. Early positron emission tomography (PET) restaging: a predictive final response in Hodgkin's disease patients. *Ann Oncol.* 2006;17:1296–1300.
31. van Tinteren H, Hoekstra OS, Smit EF, et al. Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial. *Lancet.* 2002;359:1388–1393.
32. Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. *Clin Med.* 2001;1:478–484.
33. 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.
34. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ.* 2008;336:924–926.
35. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making.* 1991;11:88–94.
36. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ.* 1996;312:71–72.
37. Muir-Gray JA. *Evidence-Based Healthcare.* Philadelphia, PA: Churchill Livingstone; 2001.
38. Kristensen FB, Adams E, Briones E, et al. Health technology assessment of PET in oncology. *Eur J Nucl Med Mol Imaging.* 2004;31:295–297.
39. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Assessing methodologic quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, Version 1.0.0*. Available at: http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/ch09_Oct09.pdf. Accessed November 3, 2011.