Journal of Nuclear Medicine, published on October 5, 2012 as doi:10.2967/jnumed.112.111427

REPLY: We welcome a scientific debate on the issue of randomized controlled trials (RCTs) in the field of PET. However, we are somewhat surprised by the tone of the letter by Drs. Hicks, Ware, and Hofman. It is up to the readers of *The Journal of Nuclear Medicine* to decide on the appropriateness of such comments as "their abstract is, at best, disingenuous and, at worst, misleading," "the authors' apparent misunderstanding of the values and principles of [evidence-based medicine (EBM)]," and "scanty assessment of the available literature." Here is our response to the main issues addressed in the letter.

In our paper we stated that RCTs could add important information to diagnostic accuracy studies in the evaluation of PET and PET/CT. Our aim was to systematically identify RCTs on PET by measuring patient-relevant outcomes in any medical indication to outline both the main fields and any gaps in research and to summarize features of study design and quality. We did not claim that the 60,162 non-RCT papers on PET we identified do not add valuable information to the body of evidence on this technology.

We agree with Hicks et al. that there has been some discussion on which trial design might be best suited to evaluate diagnostic (imaging) studies. However, there is a clear trend toward randomized designs, not only in the general methodologic literature (*1–6*) but also in nuclear medicine (*7,8*). We are not aware of any recent methodologic paper advocating test accuracy studies as the highest level of evidence in the evaluation of patient-relevant effects of diagnostic technologies. As our results indicate, increasing numbers of researchers are using RCT designs to evaluate diagnostic–therapeutic pathways involving PET. This development shows that RCTs in this field are being increasingly regarded as feasible, providing valid information on the benefits and risks of PET.

Hicks et al. state that "RCTs are most useful when the mechanism of action of treatments is not fully understood or where there is uncertainty about the benefits versus risks." In our opinion this is exactly the case in many fields in which PET is applied. They also state that "there is already abundant evidence that the diagnostic accuracy of PET/CT is superior to conventional staging approaches in many cancers." We argue that diagnostic accuracy is only a surrogate for patient-relevant outcomes (*6*). Higher diagnostic accuracy does not guarantee a benefit for patients. For example, in colon cancer staging, diagnostic accuracy studies show better accuracy for PET than for conventional imaging (*9*). However, RCTs such as that of Ruers et al. (in which the primary outcome was changed) (*10*) or PETCAM (trial NCT00265356 in the ClinicalTrials.gov registry) found no evidence of an improved diagnostic–therapeutic pathway with regard to patient-relevant outcomes. The only published RCT on PET in lymphoma found slightly more recurrences in the PET-based management arm (*11*). Therefore, more RCTs should be conducted to assess specific diagnostic–therapeutic pathways and to obtain realistic data on the benefits and harms of these strategies for patients (*6*).

Hicks et al. also state that negative biomarker trials "should not be used to justify conclusions that PET does not provide patient benefits." We fully agree and point to the seminal paper by Altman and Bland ("absence of evidence is not evidence of absence") (*12*). Hence, a nil result does not prove that PET has no benefit for patients; to prove such a benefit, prospectively planned and well-conducted RCTs with positive results are required.

In contrast to Hicks et al., we see no contradiction between our interpretation of the study of Viney et al. (*13*) and that provided by Ferrante di Ruffano et al. (*3*). We focused our appraisal on predefined methodologic characteristics that may lead to bias (see the "Data Extraction" section of our paper), which is a frequently applied approach in EBM (*14*). Ferrante di Ruffano et al. point out that even if an RCT is methodologically sound, despite a higher diagnostic accuracy other factors such as a "lack of diagnostic confidence" may prevent changes in management and lead to a nil effect. Therefore the example in the study by Viney et al. (*13*) does not conflict with our assessment of this study. In line with our point of view, Ferrante di Ruffano et al. state in their summary points that "improved accuracy is not always a necessary prerequisite for improving patient health, nor does it guarantee other downstream improvements" and "randomised controlled trials of tests can measure these processes directly to understand why and how changes to patient health have occurred."

One can always argue that RCTs with negative results are biased or not applicable to a specific clinical setting. However, as long as not a single RCT with a positive result is available for a specific clinical question, criticism of existing RCTs does not make available evidence stronger.

Hicks et al. distinguish between RCTs on "PET per se" and RCTs on "new risk-adapted therapeutic approaches," which "involve a so-called enrichment design, in which the results of PET are used to enrich the sample before randomization." In their opinion the latter "are not an evaluation of PET but rather are testing whether alternative treatment strategies can improve patient outcomes in patients stratified by PET." In times of so-called individualized medicine it no longer seems reasonable to make this distinction (*2*). In both types of studies, the diagnostic procedure as well as the treatment intervention has to be effective to achieve a patient-relevant benefit.

Hicks et al. mention that RCTs are lacking in many other areas of medicine. Is it the logical consequence of this fact to regard the conduct of RCTs in an area where RCTs are lacking as a waste of time? Do the 54 published and ongoing RCTs on PET we identified "lack clinical relevance or perspective"? There are certainly some clinical questions for which RCTs are not absolutely necessary. And Black describes specific situations in which RCTs might be difficult or impossible to conduct (*15*). None of these situations applies for the indications in which PET is usually applied, and the fact that RCTs on PET are being conducted contradicts the hypotheses of Hicks et al.

We agree with Hicks et al. that the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach "provides guidelines for using studies of test accuracy to make inferences about the likely impact on patient-important outcomes" (modeling approach). However, the GRADE authors clearly state that test accuracy is a surrogate for patient-relevant outcomes, "so studies often provide low quality evidence for recommendations about diagnostic tests even when the studies do not have serious limitations" (*6*). The GRADE authors also state that "the best way to assess any diagnostic strategy. . .is a randomised controlled trial in which investigators randomise patients to experimental or control

diagnostic approaches and measure mortality, morbidity, symptoms, and quality of life" (*6*). Moreover, high-quality-modeling approaches in the field of PET are extremely scarce.

Hicks et al. claim to demonstrate our "apparent misunderstanding of the values and principles of EBM" by criticizing the representativeness of our analysis (only 12 articles of 60,174 initially identified were included). A systematic review usually formulates a research question on the basis of patients, interventions, controls, and outcomes in a first step. In a second step a systematic search for literature on this question is performed. The more sensitive (broader) the search, the higher the number of retrieved publications. In a third step the search results are screened according to the predefined inclusion criteria (*14*). It is quite common for a systematic review to include only a relatively small number of eligible studies even though thousands of publications on a specific intervention are available in bibliographic databases. If justified, the criticism by Hicks et al. would thus apply to most of the systematic reviews ever published.

The fact that the study by Plewina et al. (*16*) included only 6 patients does not in our view undermine its "validity" as Hicks et al. point out. A crossover design as applied in this study enhances statistical power (*17*). Moreover, the fact that a significant difference was found shows that statistical power was sufficient.

Hicks et al. state that "the abstract's conclusion that a relatively high number of ongoing RCTs of PET in several oncologic fields are expected to produce robust results over the next few years" is vastly different in meaning from the statement in the body text that "it is difficult to determine whether an interaction is going to be calculated between the PET result and the effect of therapy." As noted in our paper, the above statement applies to only 9 of the 42 ongoing trials we identified, namely to those with a marker-by-treatment-interaction design. As all of the ongoing trials were identified in clinical registries, we are confident that their methodologic quality will be higher than the quality of the published trials (of which only 2 were prospectively registered). We therefore strongly disagree with the judgment of Hicks et al. that the abstract of our paper is "at best, disingenuous and, at worst, misleading."

As stated in our paper, we cannot exclude the possibility that relevant ongoing RCTs might have been overlooked with our search strategy. Moreover, because registry entries are sometimes changed, some details in our paper may no longer be up to date. Furthermore, it is sometimes difficult to identify duplicates between different registries (e.g., we have been informed that entries NCT00720070 and ISRCTN3735240 point to the same trial). However, the aim of our search in registries was to give a good estimate of the quantity of ongoing trials. It was not our aim to identify every single ongoing trial. Because we identified more than 40 ongoing trials, the fact that 1 or 2 ongoing trials might not have been identified is in our opinion negligible. However, in the case of trial NCT00882609, which Hicks et al. claim that we "failed to identify," the registry entry indicates that the primary endpoint is "an analysis of the diagnostic performance" (relative areas under the receiver-operating-characteristic curves). We prospectively defined patient-relevant outcomes as an inclusion criterion for our systematic review. We therefore excluded this study.

Hicks et al. criticize that studies in which "both the control and the treatment arms are undergoing ^{18}F-FDG PET" "cannot provide robust information about the independent contribution of PET to patient outcomes" (e.g., NCT00367341). We are convinced that it is possible to infer data on the benefits and risks of PET with this type of design (e.g., marker-by-treatment-interaction design); for methodologic details see Sargent et al. (*5*) or Lijmer and Bossuyt (*4*). For example, trial NCT00367341 states as a specific aim "to define baseline regional glucose metabolic patterns (measured using FDG PET) associated with differential clinical remission to each of two well-established, randomly delivered first-line antidepressant treatments—the [selective serotonin reuptake inhibitor] escitalopram. . .or cognitive behavioral therapy. . . ." From a statistical point of view, this aim requires PET to be handled as an effect modifier (which is usually done by calculating a statistical interaction between the PET result and the treatment effect). If an effect modification (by PET) is demonstrated in NCT00367341, it will be possible to stratify patients into those who should be treated with escitalopram and those who should receive CRT. At the same time, this result would demonstrate the benefit of PET in this specific indication (compare (*2*)).

As correctly noted by Hicks et al., trial NCT00313560 was conducted in the United States and not in Australia, as we indicated in Table 5. Nevertheless, we do not believe that such a minor error affects the "scientific rigor" of our paper.

We agree with the quote by Black that "the false conflict between those who advocate randomized trials in all situations and those who believe observational data provide sufficient evidence needs to be replaced with a mutual recognition of the complementary roles of the two approaches. Researchers should be united in their quest for scientific rigour in evaluation, regardless of the method used" (*15*). However, in the same paper, Black concludes that "after all, experimental methods depend on observational ones to generate clinical uncertainty; generate hypotheses; identify the structures, processes, and outcomes that should be measured in a trial; and help to establish the appropriate sample size for a randomised trial." In accordance with our view, observational studies are thus seen as preceding RCTs. Black also states, "when trials cannot be conducted, well designed observational methods offer an alternative to doing nothing." However, as it is feasible to conduct RCTs on PET, in our opinion researchers should no longer rely on "abundant evidence" from observational studies but focus on prospectively planned, well-conducted RCTs (with effective treatment interventions).

Van Tinteren et al. stated in 2004 that "all aspects that used to be seen as challenges to the use of randomisation for assessing medical interventions in the 1970s, such as the difficulties of performing randomised trials, their adequacy and their conclusiveness, are now being raised as arguments against randomised trials of diagnostic techniques" (*8*). This still applies (at least partly) in 2012. However, as with RCTs in drug interventions, we expect that the value of RCTs in diagnostic imaging in general and in PET in particular will be increasingly acknowledged. In the meantime we are looking forward to an objective and "evidence-based" discussion on this issue.

REFERENCES

1. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. *Med Decis Making*. 2009;29:E30–E38.
2. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, de Gramont A. Integrating biomarkers in clinical trials. *Expert Rev Mol Diagn*. 2011;11:171–182.
3. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ*. 2012;344:e686.
4. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. *J Clin Epidemiol*. 2009;62:364–373.

jnm111427-sn ■ 10/5/12

5. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol*. 2005; 23:2020–2027.
6. Schünemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008; 336:1106–1110.
7. Valk PE. Do we need randomised trials to evaluate diagnostic procedures? Against. *Eur J Nucl Med Mol Imaging*. 2004;31:132–135.
8. Van Tinteren H, Hoekstra OS, Boers M. Do we need randomised trials to evaluate diagnostic procedures? For. *Eur J Nucl Med Mol Imaging*. 2004;31: 129–131.
9. Brush J, Boyd K, Chappell F, et al. The value of FDG positron emission tomography/computerised tomography (PET/CT) in pre-operative staging of colorectal cancer: a systematic review and economic evaluation. *Health Technol Assess*. 2011;15:1–192.
10. Ruers TJ, Wiering B, van der Sijp JR, et al. Improved selection of patients for hepatic surgery of colorectal liver metastases with ^{18}F-FDG PET: a randomized study. *J Nucl Med*. 2009;50:1036–1041.
11. Picardi M, De Renzo A, Pane F, et al. Randomized comparison of consolidation radiation versus observation in bulky Hodgkin's lymphoma with post-chemotherapy negative positron emission tomography scans. *Leuk Lymphoma*. 2007;48:1721–1727.
12. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
13. 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.
14. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions, version 5.1.0. Available at: www.cochrane-handbook.org. Accessed September 27, 2012.
15. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;312:1215–1218.
16. Plewnia C, Reimold M, Najib A, Reischl G, Plontke SK, Gerloff C. Moderate therapeutic efficacy of positron emission tomography-navigated repetitive transcranial magnetic stimulation for chronic tinnitus: a randomised, controlled pilot study. *J Neurol Neurosurg Psychiatry*. 2007;78:152–156.
17. Freeman PR. The performance of the two-stage analysis of two-treatment, two-period crossover trials. *Stat Med*. 1989;8:1421–1432.

Fülöp Scheibler*
Polina Zumbé
Inger Janssen
Melanie Viebahn
Milly Schröer-Günther
Robert Grosselfinger
Elke Hausner
Stefan Sauerland
Stefan Lange
**Institute for Quality and Efficiency in Health Care*
Dillenburger Strasse 27
Cologne 51105, Germany
E-mail: fueloep.scheibler@iqwig.de

Published online ■■■■.
DOI: 10.2967/jnumed.112.111427