The Move from Accuracy Studies to Randomized Trials in PET: Current Status and Future Directions

Bettina Siepe1, Poul Flemming Hoilund-Carlsen2, Oke Gerke2,3, Wolfgang A. Weber4, Edith Motschall5, and Werner Vach6

1Department of Anesthesiology and Critical Care Medicine, Freiburg University Medical Center, Freiburg, Germany; 2Department of Nuclear Medicine, Odense University Hospital, Odense, Denmark; 3Department of Business and Economics, Centre of Health Economics Research, University of Southern Denmark, Odense, Denmark; 4Memorial Sloan-Kettering Cancer Center, New York, New York; 5Department of Medical Biometry and Medical Informatics, Freiburg University Medical Center, Freiburg, Germany; and 6Clinical Epidemiology, Department of Medical Biometry and Medical Informatics, Freiburg University Medical Center, Freiburg, Germany

Since the influential study by van Tinteren et al. published in The Lancet in 2002, there have been an increasing number of diagnostic randomized controlled trials (RCTs) investigating the benefit of PET. If they provide valid and useful information on the benefit, these studies can play an important role in informing guideline developers and policy makers. Our aim was to investigate how far the nuclear medicine community has come on its way from accuracy studies to RCTs and which issues we have to take into account in planning future studies.

Methods: We conducted a systematic review of diagnostic randomized trials, in which PET was applied in only one arm. We covered published studies as well as registered unpublished and planned studies. We considered 3 quality indicators related to the usefulness of a trial to generate evidence for a clinical benefit: use of patient-important outcome, sufficient sample size, and current standard as comparator.

Results: Fourteen published and 15 planned studies were identified. Five of the published studies and 12 of the planned studies did not use a patient-important outcome. Sample sizes were often so small that a significant result could be expected only under the assumption of a substantial reduction in the event rate. Comparators typically reflected the current standard. Conclusion: If we consider the traditional areas of primary diagnosis, staging, and follow-up, then the number and quality of RCTs on PET is currently not sufficient to provide a major source for evidence-based decisions on the clinical benefit of PET. There will also be a future need in these traditional areas to deduce the clinical benefit of PET from the results of accuracy studies. The situation may be more favorable for the areas of treatment planning and response evaluation. Choice of patient-important outcomes and sufficient sample sizes are crucial issues in planning RCTs to demonstrate the clinical benefit of using PET.

Key Words: diagnostic randomized controlled trial; patient-important outcome; positron emission tomography; systematic review

DOI: 10.2967/jnumed.113.127076

Improvement of diagnostic accuracy is no longer the ultimate criterion to establish new diagnostic procedures. Guideline developers and policy makers now require evidence of clinical benefit (1,2). Randomized controlled trials (RCTs) are often the preferred source of this evidence (3–5), because the step to deduce evidence for a clinical benefit from accuracy studies by combination with results from clinical trials or prognostic studies (1,6) is often rather cumbersome (7,8).

RCTs have gained some popularity in the area of PET since the influential multicentre study by van Tinteren et al. published in 2002 in The Lancet (9). In a randomized design, that study investigated the effect of adding 18F-FDG PET to conventional work-up before surgery in patients with suspected non–small cell lung cancer. In a recent review, Scheibler et al. (10) identified 12 published RCTs and 42 planned or unpublished RCTs from study registers. However, if RCTs are not appropriately designed to demonstrate a clinical benefit according to the established principles of evidence-based medicine and comparative effectiveness research, then they will not help overcome the limitations of accuracy studies. Scheibler et al. (10) already considered some general indicators related to the risk of bias in RCTs, such as the description of the randomization process, masking, or use of the intention-to-treat principle. In contrast, our investigation focused on aspects related to the usefulness of these studies to demonstrate a clinical benefit beyond diagnostic accuracy. Thus, we considered the use of patient-important outcomes, adequate choice of comparators, and sufficient sample size as quality indicators.

The aim of our investigation was to examine how far the nuclear medicine community has come on its way from accuracy studies to RCTs and to identify strengths and weaknesses to improve future study planning and rational promotion of clinical molecular imaging.

MATERIALS AND METHODS

Study Inclusion Criteria and Search Strategies

We searched systematically for all diagnostic RCTs in which PET was used in one arm and another modality or clinical examination was used in the other arm. No further restrictions were applied. Before the search, 2 experts from the field of nuclear medicine (PFHC, WW) each compiled a list of RCTs fulfilling the inclusion criteria based on their personal knowledge. These lists included 5 and 7 studies, respectively.

We searched in MEDLINE (National Library of Medicine; OvidSP [Ovid Technologies], 1948 to November 2011), EMBASE (Elsevier; OvidSP, 1974 to November 2011), and the Cochrane Library (Wiley...
Intercience; up to 2011). The search strategy is documented in the supplemental material (available at http://jnm.snjmjournals.org). The search was performed on November 21, 2011, and was restricted to publications since 1990. This strategy allowed us to find all RCTs identified by the experts. After the identification of eligible trials, we conducted a forward reference search (Web of Science; Thomson Reuters) on February 20, 2012. To complete the search, we also checked the references of the eligible trials.

To find unpublished or ongoing studies, we searched ClinicalTrials.gov, the International Standard Randomised Controlled Trial Number Register (ISRCTN), and the International Clinical Trials Registry platform maintained by the World Health Organization. This search was performed on January 21, 2012, using the terms diagnostic and randomized and PET (ClinicalTrials.gov) and on August 25, 2012, using the terms pet and diagnos* (ISRCTN and International Clinical Trials Registry platform).

Study Selection and Data Extraction

One reviewer (BS) screened titles and abstracts of the retrieved publications. The full text of potentially eligible publications was obtained, and a final check of the inclusion criteria was performed. In cases of multiple publications of the same trial, the primary publication was identified. The entries retrieved from the trial registers were also screened. Here, the decision about inclusion was made only on the basis of the information in the trial register. The reviewer copied the relevant passages of the identified publications or trial register entries into a spreadsheet. Study characteristics and quality criteria were then assessed jointly by BS and WV.

Study Characteristics and Quality Criteria

For each study, we extracted 2 basic study characteristics: the disease or disease stage characterizing the patients included and the clinical situation, categorized as primary diagnosis, staging, follow-up, treatment planning, or response evaluation. The specific category treatment planning refers to a situation in which the same treatment (e.g., radiation or heart surgery) was performed in both arms but PET was used in one arm to identify the regions to be treated. For registered studies, we also extracted the start of the enrollment. In addition, we evaluated the following 3 predefined criteria to assess the ability of each study to contribute to an assessment of the clinical benefit of PET imaging.

Patient-Important Outcome. If the aim of an RCT should be to assess clinical benefit, the outcome should be relevant for the patient (11). Typically, such outcomes are related to mortality, morbidity, or quality of life. They should ensure that a difference between the modalities implies an actual advantage for the patient. In diagnostic RCTs, there are 2 major sources of outcomes, which are not patient-important. The first arises from outcomes reflecting exclusively diagnostic accuracy, such as the number of correct diagnoses. As pointed out above, this is exactly what we would like to overcome in RCTs. The second stems from outcomes reflecting a management decision, the benefit of which is not known. A typical example is the frequency with which an additional, invasive diagnostic procedure is used. Reducing this frequency probably improves the quality of life on average, but we do not know whether some patients would have benefited from this procedure had it been performed. For each trial, we extracted the primary outcome and checked whether it could be regarded as being patient-important.

Sufficient Sample Size. Sufficient sample size is a crucial issue in diagnostic RCTs (12,13), as distinct improvements in diagnostic accuracy often translate into only moderate improvements in long-term outcome. For example, PET may increase both the sensitivity and the specificity from 60% to 80% (compared with a standard modality) in distinguishing between 2 different stages. If no incorrect changes occur, this implies a change in only every fifth patient. If adequate treatment at each stage implies an increase in surviving 1 y by 30 percentage points compared with inadequate treatment due to incorrect staging, then we can expect an improvement by 30 percentage points in 20% of the patients, and overall by only 6 percentage points. Regarding 1-y mortality as a binary outcome and assuming a mortality of 50% in the control arm, then we need 1,486 patients in each arm to reach 90% power in a 2-sided test with significance level 5%.

In a first step, we checked whether a sample size calculation was reported. As diagnostic RCTs can often build on results from accuracy studies, this information should be used in the sample size calculation to qualify the assumptions made. Hence, in a second step, we checked whether the assumptions made were justified by references to other publications. In a third step, we computed the relative risk, hazard ratio, or Cohen d, which is necessary to assume to obtain a power of 90% for the given trial, while taking into consideration the sample size or number of events of the study and the failure rate observed in the control arm. This post hoc power analysis gives an idea about the actual power of the study.

Current Standard as Comparator. Since the aim of RCTs should be to provide arguments for policy makers, it is important to ensure that the comparator reflects the current standard. If PET is compared in an RCT to an “old-fashioned” modality, the results cannot be used directly. To decide whether the comparator could be regarded as the current standard, we tried to identify guidelines valid at the time the study was performed and checked whether the comparator was mentioned as a possible choice.

Because of the restricted scope of information in the trial registers, in these we could apply only the criterion about the patient-important outcome.

RESULTS

Published Studies

The search in MEDLINE, EMBASE, and the Cochrane Library resulted in 3,311 citations at first and 2,228 after the duplicate check. Screening of titles and abstracts resulted in 63 potentially relevant papers. Screening of these resulted in identifying 30 papers describing 14 trials that fulfilled our inclusion criterion. Table 1 summarizes the results of our investigation for the 14 published studies identified. Ten studies were from oncology, 3 from cardiology, and 1 from neurology. Six studies considered staging, 5 treatment planning, and 3 follow-up evaluation. One publication (14) reported a study protocol, not a finished study. The study of Plewnia et al. (15) used a cross-over design.

In 5 of the 14 studies, the outcome was not a patient-important one. The number of positive scans used by Mullani et al. (16) covered both false-positive and true-positive scans, so that any increase might have been due to just an increasing number of false-positive scans, which is not an advantage for the patients. Correct upstaging used by Maziak et al. (17) implies avoiding unnecessary surgery and, hence, is an advantage, but its frequency has to be balanced against the frequency of incorrect upstaging, which implies that patients miss the advantage of surgery. Indeed, this study observed a significant increase in both correct upstaging by PET and incorrect upstaging by PET. Recurrence during follow-up was the primary outcome in the study of Sobhani et al. (18). In treatment trials, this is a patient-important outcome reflecting morbidity. However, Sobhani et al. did not use PET as a means to prolong the true time until recurrence but rather tested whether PET could shorten the time until detection of recurrences, which is a diagnostic decision and not a therapeutic effect. Actually, the authors reported that only 27 of the 44 recurrences they observed...
could be confirmed by biopsy or surgery. Therefore, since it remains unclear whether their patients actually benefited from the earlier detection, this is not a patient-important outcome either. Similar arguments apply to one further study (19). Finally, one study (20) considered the number of investigations needed to finalize staging as primary outcome. Reducing this number may be beneficial to the quality of life of the patient. However, this advantage may be counterbalanced by a poorer quality of staging. Consequently, this outcome can be regarded as patient-important only if we are sure that there was no loss in diagnostic accuracy. The authors provided additional information on the agreement between clinical and final staging in both arms addressing this issue. However, with limited sample sizes it is difficult to ensure true equivalence.

Nine studies used a patient-important outcome. Three of these studies (21–23) used long-term outcomes related to (disease-free) survival. One study (15) applied a disease-specific quality-of-life measure. The outcome used in the study of Viney et al. (24) was the performance of a thoracotomy. As that is a management decision, we have to be careful: if a diagnostic method never points to performing a thoracotomy, this method may turn out to be the best one. However, the authors listed in their Table 3 an overview of the reasons for not performing a thoracotomy, and in the text they provided an explicit reason why the thoracotomy would be futile for each patient lacking thoracotomy. Hence, in this case, to regard the outcome as being patient-important seems justified, as it reflects avoidance of an unnecessary invasive procedure. Another...

<table>
<thead>
<tr>
<th>Study</th>
<th>Disease</th>
<th>Clinical situation</th>
<th>Primary outcome</th>
<th>Patient-important outcome</th>
<th>Comparator is current standard</th>
<th>Sample size calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beanlands (2007)</td>
<td>Coronary heart disease</td>
<td>Treatment planning</td>
<td>Composite clinical endpoint at 1 y</td>
<td>Yes</td>
<td>Yes (33)</td>
<td>Yes</td>
</tr>
<tr>
<td>de Bree (2007)</td>
<td>Laryngeal carcinoma</td>
<td>Follow-up</td>
<td>Futile indication for direct laryngoscopy</td>
<td>(Yes)</td>
<td>NA</td>
<td>Yes</td>
</tr>
<tr>
<td>Fischer (2007)</td>
<td>NSCLC</td>
<td>Staging</td>
<td>Futile thoracotomy</td>
<td>Yes</td>
<td>Yes (34,35)</td>
<td>No</td>
</tr>
<tr>
<td>Herder (2007)</td>
<td>NSCLC</td>
<td>Staging</td>
<td>Number of diagnostic investigations</td>
<td>(No)</td>
<td>(Yes)* (36)</td>
<td>Yes</td>
</tr>
<tr>
<td>Maziak (2009)</td>
<td>NSCLC</td>
<td>Staging</td>
<td>Correct upstaging of cancer</td>
<td>No</td>
<td>Yes (34,35)</td>
<td>For one arm</td>
</tr>
<tr>
<td>Monteil (2010)</td>
<td>NSCLC</td>
<td>Follow-up</td>
<td>Recurrence or new tumor detected</td>
<td>No</td>
<td>Yes (36)</td>
<td>No</td>
</tr>
<tr>
<td>Plewnia (2007)</td>
<td>Tinnitus</td>
<td>Treatment planning</td>
<td>Tinnitus distress measure</td>
<td>Yes</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>Ruers (2009)</td>
<td>Colorectal liver metastases</td>
<td>Staging</td>
<td>Futile laparotomy</td>
<td>Yes</td>
<td>(Yes) (37)</td>
<td>Yes</td>
</tr>
<tr>
<td>Siebelink (2001)</td>
<td>Coronary heart disease</td>
<td>Treatment planning</td>
<td>Cardiac event-free survival</td>
<td>Yes</td>
<td>?</td>
<td>For one arm</td>
</tr>
<tr>
<td>Sobhani (2008)</td>
<td>Colorectal cancer</td>
<td>Follow-up</td>
<td>Recurrence after 9 and 15 mo of follow-up</td>
<td>No</td>
<td>Yes (38,39)</td>
<td>Yes</td>
</tr>
<tr>
<td>Tsai (2010)</td>
<td>Cervical cancer</td>
<td>Treatment planning</td>
<td>Therapeutic outcomes (survival)</td>
<td>Yes</td>
<td>Yes (40)</td>
<td>No</td>
</tr>
<tr>
<td>van Tinteren (2002)</td>
<td>NSCLC</td>
<td>Staging</td>
<td>Futile thoracotomy</td>
<td>Yes</td>
<td>Yes (36)</td>
<td>For both arms</td>
</tr>
<tr>
<td>Viney (2004)</td>
<td>NSCLC</td>
<td>Staging</td>
<td>Thoracotomy avoided</td>
<td>Yes</td>
<td>Yes (36)</td>
<td>No</td>
</tr>
</tbody>
</table>

*Description was too incomplete for comparison with international guidelines, but authors claimed that it was “according to international guidelines.”
NA = not applicable; NSCLS = non–small cell lung cancer; ? = comparator could not be evaluated.
Parentheses around the words yes or no indicate an uncertain assessment.
Table 2

Effects to Be Assumed in Each Study to Reach Power of 90%

<table>
<thead>
<tr>
<th>Study</th>
<th>Binary outcome*</th>
<th>Time-to-event outcome†</th>
<th>Continuous outcome‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Failure rate control arm</td>
<td>n</td>
<td>RR (90%)</td>
</tr>
<tr>
<td>de Bree (14)</td>
<td>0.38</td>
<td>75/75</td>
<td>0.36</td>
</tr>
<tr>
<td>Fischer (25)</td>
<td>0.42</td>
<td>91/98</td>
<td>0.46</td>
</tr>
<tr>
<td>Montel (19)</td>
<td>0.72</td>
<td>33/36</td>
<td>0.44</td>
</tr>
<tr>
<td>Mullani (16)</td>
<td>0.71</td>
<td>105/105</td>
<td>0.68</td>
</tr>
<tr>
<td>Maziak (17)</td>
<td>0.93</td>
<td>162/167</td>
<td>0.86</td>
</tr>
<tr>
<td>Ruers (26)</td>
<td>0.45</td>
<td>75/75</td>
<td>0.43</td>
</tr>
<tr>
<td>Viney (24)</td>
<td>0.98</td>
<td>92/91</td>
<td>0.85</td>
</tr>
<tr>
<td>van Tinteren (9)</td>
<td>0.41</td>
<td>96/92</td>
<td>0.46</td>
</tr>
</tbody>
</table>

*Data are failure rate in control arm, number of patients in control arm and number in PET arm, and relative risk (RR) to be assumed.
†Data are number of events observed and number in overall sample, and hazard ratio (HR) to be assumed.
‡Data are number of patients in control arm and number in PET arm, and Cohen d to be assumed. One study with crossover design (17) is not included in this table.

A sample size calculation was performed for most studies. However, only 2 studies succeeded in presenting explicit justification for the assumed effects in both arms. Moreover, if we look at the effects to be assumed to reach a power of 90% (Table 2), we observe that in most of these studies relative risks or hazard ratios of less than 0.5 have to be assumed to reach this power. In treatment research, such large effects are rare, and Djulbegovic et al. (27) suggested regarding the few studies that were able to demonstrate effects of this magnitude as “breakthroughs.” On the other hand, all 3 studies using futile surgery as primary outcome (9,25,26) actually estimated relative risks in the magnitude of 0.5. A Cohen d of 0.3 is typically regarded as a rather moderate effect.

With respect to the choice of the comparator, we could identify appropriate guidelines for 9 studies from oncology (Table 1). In all cases, the comparator chosen was not in conflict with the existing guidelines. In one study (26), CT was applied in all patients of the control arm, although this was not yet recommended in the guidelines. For 2 cardiologic studies, the onset of the study was not sufficiently specified to locate the corresponding guideline. The comparator in the study of Beanlands et al. (21) was in accordance with existing guidelines.

Registered Studies

The search in the trial registers resulted in 263 studies (clinicaltrials.gov, 63; International Clinical Trials Registry Platform, 166; ISRCTN, 34). Removing duplicates and screening the information available in the trial register resulted in 20 eligible trials. Among these, 5 were already included in the 14 published studies. The remaining 15 studies are summarized in Table 3.

A sample size calculation was performed for most studies. However, only 2 studies succeeded in presenting explicit justification for the assumed effects in both arms. Moreover, if we look at the effects to be assumed to reach a power of 90% (Table 2), we observe that in most of these studies relative risks or hazard ratios of less than 0.5 have to be assumed to reach this power. In treatment research, such large effects are rare, and Djulbegovic et al. (27) suggested regarding the few studies that were able to demonstrate effects of this magnitude as “breakthroughs.” On the other hand, all 3 studies using futile surgery as primary outcome (9,25,26) actually estimated relative risks in the magnitude of 0.5. A Cohen d of 0.3 is typically regarded as a rather moderate effect.

With respect to the choice of the comparator, we could identify appropriate guidelines for 9 studies from oncology (Table 1). In all cases, the comparator chosen was not in conflict with the existing guidelines. In one study (26), CT was applied in all patients of the control arm, although this was not yet recommended in the guidelines. For 2 cardiologic studies, the onset of the study was not sufficiently specified to locate the corresponding guideline. The comparator in the study of Beanlands et al. (21) was in accordance with existing guidelines.

Registered Studies

The search in the trial registers resulted in 263 studies (clinicaltrials.gov, 63; International Clinical Trials Registry Platform, 166; ISRCTN, 34). Removing duplicates and screening the information available in the trial register resulted in 20 eligible trials. Among these, 5 were already included in the 14 published studies. The remaining 15 studies are summarized in Table 3.

Fourteen studies are from oncology, and one from cardiology. In these studies the clinical situation was treatment planning for 5 studies, staging for 4, primary diagnosis for 2, follow-up for 2, and response evaluation for 1. In most studies, the primary outcome was related to accuracy, diagnostic decisions, or management decisions and, therefore, could not be regarded as patient-important. Only 3 studies used an unequivocally patient-important outcome. For 2 further studies, this decision depends on whether the events considered will be validated. One study used the response rate, which is today typically regarded as a surrogate outcome and not a true patient-important outcome (28). However, the description of the primary outcome was often close to a description of the general aim of the study, and hence we cannot exclude that the primary outcome actually chosen in the analysis will be patient-important.

Three clinical situations contributed at least 2 studies both among the published studies and among the registered studies: staging, treatment planning, and follow-up. The median sample sizes in the published studies were 188, 210, and 130, respectively. In the registered studies, the corresponding median sample sizes were 220, 288, and 214, suggesting a slight increase.

Comparison with Results of Scheibler et al.

The review of Scheibler et al. (10) identified 12 published studies. We identified 3 additional studies that were probably excluded by the authors for the following reasons: use of 18F-FDG coincidence imaging by dual-head camera rather than a real PET scanner (19), focus on risk factors instead of modality comparison (16), and the
nature of a study protocol (14). One study included in the review of Scheibler et al. was excluded in our review, as PET was administered to all patients (29). Scheibler et al. identified 42 registered studies, in contrast to 20 identified by us, 4 of which were not included in their review. The much larger number of Scheibler et al. probably resulted from their searching for all randomized trials, whereas we included the term \textit{diagnos*} in the search strategy. Moreover, we included only comparative RCTs, comparing PET with a competitor. Of the 28 additionally identified studies, 18 used an enrichment or interaction design; that is, they were not comparative in our sense. Six studies considered treatment planning, and 10 studies considered response evaluation. If we focus on primary diagnosis, staging, and follow-up evaluation—the most common context of accuracy studies—there is only one additional comparative study (ClinicalTrials.gov identifier NCT00329706) in the review of Scheibler et al. This study compared the impact of making the result of a PET scan available to the clinician immediately or after 2 y. This study actually was identified in our search but was excluded, because both arms were PET-based.

**DISCUSSION**

The systematic reviews by Scheibler et al. (10) and us both indicate an increasing number of RCTs on PET in recent years. One nice example illustrating this move from accuracy studies to RCTs is the increasing number of trials comparing PET-based and

### TABLE 3

Selected Characteristics of Studies Found in Clinical Trial Registers

| Register ID | Study register Acronym Disease Clinical situation Primary important primary outcome First enrollment Planned overall sample size |
|-------------|------------------------------------------------------------------|---------------------------------------------------------------|-----------------|-----------------|-----------------|
| NCT00136864 | WHO/ct.gov PET-START NSCLC Treatment planning Upstaging No Aug 2004 400 |
| NCT0082609* | ct.gov Cancer (breast, prostate, lung) Unclear Diagnostic performance No Jan 2009 550 |
| NCT00265356 | WHO/ct.gov PETFAM CRC liver metastases Staging/ treatment planning Change in management No Nov 2005 404 |
| NCT00976053* | ct.gov CAD Diagnostic (known CAD) Treatment planning Diagnostic failure Treatment delivered No June 2009 330 |
| NCT00895349 | ct.gov PET LACE Cervical cancer Primary diagnosis Cancer diagnosed No Apr 2010 288 |
| NCT00169598 | ct.gov TEPELY Lymphoma (HD, non-HD) Diagnostic performance Therapeutic prescription No Feb 2002 80 |
| NCT00199654 | ct.gov CRC Follow-up Time to CRC relapse (Yes) Feb 2004 376 |
| NCT01490026* | ct.gov CUP Project CUP Staging Detection of primary tumor possible No Nov 2011 220 |
| NCT00954148* | ct.gov Cancer Follow-up 5-y survival, cost and time to identification of new disease Yes Sep 2009 53 |
| NCT0117023 | ct.gov NSCLC Treatment planning Change in response rate (Yes) Sep 2008 100 |
| NCT00720070 | ct.gov PET/CT Head and neck cancer Staging Overall survival Yes Sep 2007 560 |
| NCT00433433 | ct.gov Hodgkin lymphoma Prostate cancer Early response evaluation Progression-free survival Change in management No Oct 2006 1,787 |
| ACTRN12608000641392 | WHO Prostate cancer Treatment planning Change in management No Oct 2008 100 |
| ISRCTN49573946 | ISRCTN BOOST Lung cancer Primary diagnosis/ staging Time to treatment decision No April 2008 168 |

*Studies not included in review by Scheibler et al. (10).

WHO = International Clinical Trials Registry (World Health Organization); ct.gov = clinicaltrials.gov; NSCLC = non–small cell lung cancer; CAD = coronary artery disease; HD = Hodgkin disease; CRC = colorectal cancer; CUP = cancer of unknown primary. Parentheses around the words \textit{yes} or \textit{no} indicate an uncertain assessment.
non–PET-based treatment planning. In this area, accuracy studies can typically provide information on improved sensitivity or specificity only on the level of single lesions or local regions. Whether such an improvement results in better treatment planning cannot be directly addressed in these studies because this result will depend on the specific strategy chosen and on the accuracy at the patient level, for example, whether all lesions have been detected. Interestingly, we overlooked some of these planned studies, as the study registration did not mention anything about diagnosing, probably because most principal investigators regard such studies as having a therapeutic rather than a diagnostic focus. Another area with an increasing number of RCTs is response evaluation. This increase may reflect the fact that response evaluation is often a rather new topic with no established methods that can serve as a gold standard in accuracy studies. Prognostic studies using survival as a gold standard are hampered by the fact that effective second-line therapies invalidate this gold standard (8). Consequently, a direct comparison of no response evaluation against a combination of response evaluation by PET and an effective second-line therapy is often the only way to demonstrate the value of PET.

The picture is less favorable if we focus on those clinical situations in which accuracy studies are feasible and have a long tradition: primary diagnosis, staging, and follow-up evaluation. These are also areas in which PET competes with existing modalities. PET has been considered a diagnostic modality for nearly all cancer entities, typically for various different clinical situations (3,30). The 10 published studies (5 alone on staging in non–small cell lung cancer patients) and 6 planned studies identified in our review and fitting into these areas have, then, to be regarded as a small number. This may simply reflect the fact that it is easier and cheaper to aim to demonstrate an improvement in accuracy in a single-center study than it is to aim to translate a better accuracy into a pathophysiologically more correct and, consequently, more patient–favorable treatment in a multidisciplinary clinical study. It is time that in the field of nuclear imaging, also, we overcome the traditional distinction between clinical trials and diagnostic studies, similar to the case in biomarker research (31,32). Close collaboration between clinical departments and imaging units and financial support by funding agencies are important issues in reaching this aim.

However, the impact of RCTs may not be restricted only by their small number. According to our review, there are still many RCTs using accuracy or management decisions as primary outcome, that is, RCTs that fail to use patient–important outcomes. Such is the particular case among the clinical settings mentioned above, in which half the published RCTs do not use a patient–important outcome. These studies are of limited value because guidelines and reimbursement decisions should be based on evidence for a clinical benefit. Our post hoc power analysis of the published RCTs also indicates a lack of power to demonstrate clinically relevant effects that are not at the level of a “breakthrough.” Unfortunately, according to our review the intended sample sizes in the planned studies did not differ substantially from those observed in the published studies, and they do not plan to use more efficient designs such as randomizing only discordant patients. Consequently, this problem may persist. In contrast, the choice of the comparator was adequate for all studies as far as we could judge.

Besides the 3 quality indicators considered in this paper, further aspects are important to allow a generalization from the RCT to clinical practice. These include a description of basic technical features of the imaging procedure, a clear definition of the target population, and a description of the actual recruitment setting. For all 3 aspects, we could observe deficits in the reporting of some trials (supplemental data).

The choice of a patient–important outcome is a prerequisite for a convincing assessment of the clinical benefit of PET. However, the fact that this is not a trivial task may explain why we have few RCTs covering the above-mentioned clinical situations. In our review, we could actually observe only 3 different types of patient–important outcomes: futile surgery, (disease-free) survival, and quality of life. Futile surgery is special not only because it is restricted to patient courses with a surgical treatment option but also because, as an outcome, it needs a valid assessment of actually nonoperated surgeries to be futile, that is, a gold standard applicable in the absence of surgery. The absence of such a gold standard is a major limitation in many accuracy studies and, hence, is also the case for RCTs. It is somewhat suspicious that in all 3 studies with futile surgery as outcome (9,25,26), no single case of a false-negative decision—a wrong suggestion to avoid surgery—appeared, despite the overall reduction in the number of surgeries from 226 to 190 when the control arms were compared with the PET arms. A consensus is needed on which outcomes should be regarded as patient–important in diagnostic RCTs. Moreover, to inform reimbursement decisions adequately, costs should also be included as outcomes to allow cost-effectiveness analyses.

CONCLUSION

Although the number of RCTs in PET is increasing, RCTs are still lacking in those areas in which accuracy studies have traditionally been performed. Thus, in PET research the move from accuracy studies to RCTs seems to be rather slow, suggesting that in the near future we will still have to deduce clinical benefit from accuracy studies to inform guidelines or reimbursement decisions. Furthermore, in the future the emphasis should not only be on randomization when increasing the number of RCTs but also on quality criteria ensuring that these studies can actually contribute to an assessment of a genuine clinical benefit. To achieve these improvements, the choice of patient–important outcomes and sufficient sample sizes may be the main issues.

DISCLOSURE

The costs of publication of this article were defrayed in part by the payment of page charges. Therefore, and solely to indicate this fact, this article is hereby marked “advertisement” in accordance with 18 USC section 1734. No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We acknowledge the excellent support of Franklin Torres in editing and proofreading the manuscript and of Monika Richards in preparing the final version. We are grateful to the 3 reviewers for their valuable comments.

REFERENCES

FROM ACCURACY STUDIES TO RCTS • Siepe et al. 7


