Positron Emission Tomography in Staging of Esophageal Cancer
Westreenen, Henderik Leendert van

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2005

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
ESOPHAGEAL CANCER: CT, EUS, AND FDG-PET FOR ASSESSMENT OF RESPONSE TO NEOADJUVANT THERAPY - SYSTEMATIC REVIEW

M Westerterp¹
HL van Westreenen²
JH Reitsma³
OS Hoekstra⁴
PL Jager⁵
J Stoker⁶
P Fockens⁷
BL van Eck-Smit⁸
JThM Plukker²
JJB van Lanschot¹
GW Sloof⁸

Departments of Surgery¹, Clinical Epidemiology and Biostatistics³, Radiology⁵, Gastroenterology⁷, Nuclear Medicine⁸
Academic Medical Center, Amsterdam, The Netherlands
Department of Nuclear Medicine⁴, VU Medical Center, Amsterdam, The Netherlands
Departments of Surgery¹, Nuclear Medicine/PET center⁵
University Hospital Groningen, The Netherlands

Radiology 2005;236;841-851
ABSTRACT

Introduction: To compare diagnostic accuracy of computed tomography (CT), endoscopic ultrasonography (EUS), and \(^{18}\)F-fluorodeoxyglucose positron emission tomography (FDG-PET) for assessment of response to neoadjuvant therapy in patients with esophageal cancer by using a systematic review of the literature.

Materials and Methods: Medline and Embase databases and Cochrane Database of Systematic Reviews were searched for relevant studies. Two reviewers independently assessed the methodological quality of each study. Summary receiver operating characteristic (ROC) analysis was used to summarize and compare the diagnostic accuracy of the three modalities.

Results: Four studies with CT, 13 with EUS, and seven with FDG-PET met inclusion criteria. Percentages of maximum score in regard to methodological quality ranged from 15% to 100%. Summary ROC analysis could be performed for three studies with CT, four with EUS, and four with FDG-PET. The maximum joint values for sensitivity and specificity were 54% for CT, 86% for EUS, and 85% for FDG-PET. Accuracy of CT was significantly lower than that of FDG-PET (\(P < 0.006\)) and of EUS (\(P < 0.003\)). Accuracy of FDG-PET and that of EUS were similar (\(P = 0.839\)). In all patients, CT was always feasible, whereas EUS was not feasible in 6% of the patients, and FDG-PET was not feasible in less than 1% of the patients.

Conclusion: CT has poor accuracy for assessment of response to neoadjuvant therapy in patients with esophageal cancer. EUS and FDG-PET have equivalent good accuracy, but EUS is not always feasible after chemotherapy and radiation therapy. FDG-PET seems to be a promising noninvasive tool for assessment of neoadjuvant therapy in patients with esophageal cancer.
INTRODUCTION

Esophageal cancer (EC) has an unfavorable prognosis among digestive tract malignancies.\textsuperscript{1,2,3} The best option for curative treatment for patients with esophageal cancer is radical surgery\textsuperscript{4}, with a long-term survival of only 25%.\textsuperscript{5,6} Therefore, various forms of neoadjuvant or adjuvant multimodality therapy have been evaluated in an effort to improve these survival results. Neoadjuvant therapy is aimed at the eradication of lymphatic and/or hematogenous micrometastases and metastases, with improvement of survival, and at shrinkage of the primary tumor, with an improved radical resectability rate. At many institutions, neoadjuvant chemotherapy and radiation therapy is applied to improve long-term outcome, especially after the recent publication of favorable long-term results of a randomized trial from the Medical Research Council Oesophageal Cancer Working Group, in which neoadjuvant chemotherapy followed by surgery was compared with surgery alone.\textsuperscript{7}

In a large proportion of patients, however, an insufficient objective response is achieved. These patients do not benefit from neoadjuvant therapy but suffer from toxic side effects while appropriate surgical therapy is delayed. For this reason, a diagnostic test that can accurately predict tumor response early in the course of neoadjuvant therapy is of crucial importance. Currently, there is no universally accepted, reproducible, and reliable means of monitoring the response of esophageal cancer to chemotherapy. Response to therapy currently is evaluated by using morphological imaging, such as computed tomography (CT), and endoscopic ultrasonography (EUS).\textsuperscript{8-10} General restrictions of these methods are the difficulty in distinguishing viable tumor from necrotic or fibrotic tissue and the delay between cell kill and tumor shrinkage.\textsuperscript{6,11,12}

With \textsuperscript{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET), alterations in tissue metabolism that generally precede anatomic change are reflected. FDG preferentially accumulates in cells with high rates of glucose utilization (e.g. the brain, the myocardium, and most solid malignancies). The accumulation of FDG in tumor cells can be measured noninvasively by using PET. There are various approaches for analytical methods ranging from visual assessment (qualitative) to semiquantitative indices (e.g. standardized uptake value (SUV)). FDG-PET has been shown to be sensitive in the imaging of several malignancies (e.g. lymphoma, lung cancer, colorectal cancer, and head and neck cancer) and has been shown to be promising for detection of the response to nonsurgical therapy in breast cancer.\textsuperscript{13} Thus, the purpose of our study was to compare the diagnostic accuracy of CT, EUS, and FDG-PET for the assessment of the response to neoadjuvant therapy in patients with esophageal cancer by using a systematic review of the literature.

MATERIALS AND METHODS

Search strategy and inclusion criteria

Two authors (MW and HLVW) independently performed a formal computer-assisted search of the medical databases Medline (January 1980 to January 2004), Embase (January 1980 to January 2004), and for the Cochrane Database of Systematic Reviews (January 1980 to January 2004). The following keywords, including comparable synonyms, and medical subject heading were used: ‘positron emission tomography’, ‘computed tomography’,
'endosonography', 'esophageal cancer' and 'neoadjuvant therapy OR response' without any language restrictions.

A manual search with cross-reference of the eligible articles was performed to identify additional relevant articles. We did not include conference abstracts because of the limited data presented in them.

The same two authors independently assessed articles for possible inclusion in the review by checking titles and abstracts. Included were clinical studies that fulfilled all the following inclusion criteria: evaluation of CT, EUS or FDG-PET in the assessment of the response to neoadjuvant therapy (before and after therapy); histologic proof of cancer of the esophagus, gastroesophageal junction or gastric cardia; use of a valid reference standard (i.e. pathologic findings); and a number of patients of 10 or greater. Duplicate studies involving the same patients were excluded. The final decision about inclusion was based on the full article. Disagreement was resolved in a consensus meeting.

Quality assessment

The same two reviewers who performed the search and assessed publications for inclusion independently assessed the methodological quality of the included studies. They used the list of criteria recommended by the Cochrane Methods Working Group on Systematic Review of Screening and Diagnostic Tests, with some modifications.14 The complete list of criteria is shown in Table 1.

Internal validity criteria were assigned a score of positive (adequate methods), negative (inadequate methods, potential for bias), or unclear (if insufficient information had
been provided about a specific item). External validity criteria were assessed to evaluate generalizability. The criteria for external validity were assigned a positive score if sufficient information was provided to judge generalizibility of findings. Description of the index test was assigned a positive score when sufficient details were described to permit replication of the test. After the consensus meeting, it was decided to change unclear scores to negative scores. Quality scores were expressed as a percentage of the maximum score. Subtotals were calculated for internal validity (maximum score, seven) and external validity (maximum score, six) criteria separately.

**Data and statistical analysis**

For all included studies, an attempt was made to extract the 2×2 table that was created with cross-classification of the patients based on the results of CT, EUS, and FDG-PET and the final outcome that was confirmed with the reference standard (histologic findings). From these tables, we calculated sensitivity and specificity, together with exact 95% confidence intervals, on the basis of the binomial distribution for each study. Forrest plots were used to present data. For each study, we plotted the pairs of sensitivity and specificity in receiver operating characteristic (ROC) space and used the approach of Moses et al., Littenberg and Moses, and Irwig et al to summarize the data by fitting the summary ROC curve.\(^{15,16}\) This method models test accuracy, defined by the log diagnostic odds ratio (D), as a linear function of the test threshold (S). S represents the (implicit) threshold for a positive test result for each study. The model \(D = a + bS\) was used to capture the variation in diagnostic odds ratio caused by differences in test threshold between studies. This model was fitted using equally weighted least squares regression because this produces results that are more consistent with a random-effects approach and allows additional variance in accuracy beyond the test threshold and sampling error.\(^{17}\) We transformed the regression line back to the original ROC space to obtain the summary ROC curve. Because the log odds ratio is difficult to interpret clinically, we expressed the results in terms of the maximum joint sensitivity and specificity, also known as the Q-point. This is the point on the summary ROC curve closest to the optimal upper left corner of the ROC plot. It is also the point where sensitivity equals specificity.

We used the Q-point to test for a difference in accuracy between techniques. This test was directly derived from the model by comparing the diagnostic odds ratio at \(S=0\) among the three modalities. For clinical purposes, however, a high negative predictive value of a modality is necessary to avoid erroneous discontinuation of neoadjuvant therapy. Therefore, the predicted value of specificity at a high value of sensitivity of 90% was calculated from the summary ROC regression line. A difference with a two-sided P value less than 0.05 was considered significant. To avoid division by zero in the calculation of the diagnostic odds ratio, the standard correction of adding 0.5 to all four cells of the 2×2 table was applied when one of the four cells contained a zero. Statistical analyses were performed by using a statistical software system (SAS Institute Inc., version 8.02, Cary, NC, USA).
Table 2. Characteristics of the Included Studies

<table>
<thead>
<tr>
<th>Modality</th>
<th>N</th>
<th>Excluded</th>
<th>M/F</th>
<th>Age (range)</th>
<th>Histology (AC/SCC/other)</th>
<th>Therapy</th>
<th>Stage of disease</th>
<th>Modality response</th>
<th>PA response</th>
<th>Prevalence responders</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walker 1991</td>
<td>38</td>
<td>-</td>
<td>32/6</td>
<td>45-73</td>
<td>19/15/4</td>
<td>CT</td>
<td>I-III</td>
<td>Miller et al</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Griffith 1999</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>45/-/-</td>
<td>CRT</td>
<td>II-IV</td>
<td>MCSA</td>
<td>Mandard</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>Jones 1999</td>
<td>50</td>
<td>-</td>
<td>34/16</td>
<td>43-81</td>
<td>CRT</td>
<td>I-IIB</td>
<td>ECOG</td>
<td>Other</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td>Kroep 2003</td>
<td>13</td>
<td>2</td>
<td>12/1</td>
<td>50-69</td>
<td>CT</td>
<td>III</td>
<td>WHO</td>
<td>Mandard</td>
<td>36%</td>
<td></td>
</tr>
<tr>
<td>EUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rice 1992</td>
<td>13</td>
<td>2</td>
<td>13</td>
<td>50-69</td>
<td>13/-/-</td>
<td>CT</td>
<td>IIA-IV</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
</tr>
<tr>
<td>Hordijk 1993</td>
<td>11</td>
<td>1</td>
<td>7/4</td>
<td>55-75</td>
<td>CRT</td>
<td>I-III</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dittler 1994</td>
<td>18</td>
<td>-</td>
<td>-</td>
<td></td>
<td>CRT</td>
<td>III</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isenberg 1998</td>
<td>31</td>
<td>8</td>
<td>22/9</td>
<td>62^§</td>
<td>CRT</td>
<td>II-IIB</td>
<td>Restaging MCSA</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowrey 1999</td>
<td>17</td>
<td>-</td>
<td>10/7</td>
<td>47-72</td>
<td>CRT</td>
<td>II-IIB</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laterza 1999</td>
<td>87</td>
<td>25</td>
<td>-</td>
<td>-/-7/-</td>
<td>CRT</td>
<td>I-IIB</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zuccaro 2000</td>
<td>59</td>
<td>-</td>
<td>30-77</td>
<td>41/18/-</td>
<td>CRT</td>
<td>I-IIB</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beseth 2000</td>
<td>26</td>
<td>6</td>
<td>24/2</td>
<td>38-74</td>
<td>CRT</td>
<td>I-IIB</td>
<td>Restaging</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chak 2000</td>
<td>59</td>
<td>-</td>
<td>44/15</td>
<td>61^#</td>
<td>CRT</td>
<td>II-IIB</td>
<td>MCSA</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PET</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arslan 2002</td>
<td>24</td>
<td>-</td>
<td>24/0</td>
<td>36-82</td>
<td>CRT</td>
<td>I-IIIB</td>
<td>SUV</td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kato 2002</td>
<td>10</td>
<td>-</td>
<td>8/2</td>
<td>36-77</td>
<td>CRT</td>
<td>III</td>
<td>SUV</td>
<td>JSED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downey 2003</td>
<td>39</td>
<td>22</td>
<td>34/5</td>
<td>36-76</td>
<td>CRT</td>
<td>I-IIIB</td>
<td>SUV</td>
<td>Not defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brucher 2001</td>
<td>27</td>
<td>3</td>
<td>23/4</td>
<td>38-61</td>
<td>CRT</td>
<td>II-IIIB</td>
<td>SUV</td>
<td>Mandard</td>
<td>54%</td>
<td></td>
</tr>
<tr>
<td>Weber 2001</td>
<td>40</td>
<td>3</td>
<td>37/3</td>
<td>44-66</td>
<td>CRT</td>
<td>III</td>
<td>SUV</td>
<td>Mandard</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td>Flamen 2002</td>
<td>36</td>
<td>-</td>
<td>28/6</td>
<td>60^</td>
<td>CRT</td>
<td>II-IIIB</td>
<td>Delta TUR</td>
<td>Other</td>
<td>39%</td>
<td></td>
</tr>
<tr>
<td>Kroep 2003</td>
<td>13</td>
<td>3</td>
<td>12/1</td>
<td>50-69</td>
<td>CT</td>
<td>III</td>
<td>SUV/SKM/NLR</td>
<td>Mandard</td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

N: number of patients included; † number of excluded patients; M/F: male/female; AC/SCC: adenocarcinoma/squamous cell carcinoma; PA: pathology; CT: chemotherapy; CRT: chemoradiotherapy; MCSA: maximal cross-sectional area; ECOG: Eastern Cooperative Oncology Group solid tumor response criteria; WHO: World Health Organisation criteria; JSED: Japanese Society for Esophageal Diseases; SUV: standardized uptake value; delta-TUR: tumor to liver uptake ratio; SKM: simplified kinetic method; NLR: non linear regression; #: median; § mean
Table 3. Quality Assessment of the Diagnostic Studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>CT Internal Validity</th>
<th>CT External Validity</th>
<th>Total IV</th>
<th>Total EV</th>
<th>% of Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walker</td>
<td>1991</td>
<td>+ + - - + + + - - -</td>
<td>+ + + + + + + + + + +</td>
<td>4</td>
<td>2</td>
<td>46%</td>
</tr>
<tr>
<td>Griffith</td>
<td>1999</td>
<td>+ + + - + - + + - +</td>
<td>+ + - - - + - + + + +</td>
<td>5</td>
<td>1</td>
<td>46%</td>
</tr>
<tr>
<td>Jones</td>
<td>1999</td>
<td>+ + + + + + + + + + +</td>
<td>+ + + + + + + + + + +</td>
<td>6</td>
<td>6</td>
<td>92%</td>
</tr>
<tr>
<td>Kroep</td>
<td>2003</td>
<td>+ + - + + - + + + + +</td>
<td>+ + + + + + + + + + +</td>
<td>5</td>
<td>5</td>
<td>77%</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>62%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>EUS Internal Validity</th>
<th>EUS External Validity</th>
<th>Total IV</th>
<th>Total EV</th>
<th>% of Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice</td>
<td>1992</td>
<td>- - + - - + + - + -</td>
<td>+ + + - - + - + + +</td>
<td>3</td>
<td>4</td>
<td>54%</td>
</tr>
<tr>
<td>Hordijk</td>
<td>1993</td>
<td>- - + - - + + + - +</td>
<td>+ + + - + + + + + +</td>
<td>3</td>
<td>5</td>
<td>62%</td>
</tr>
<tr>
<td>Dittler</td>
<td>1994</td>
<td>- - - - - + + - + +</td>
<td>+ + + + + + + + + +</td>
<td>1</td>
<td>1</td>
<td>15%</td>
</tr>
<tr>
<td>Giovannini</td>
<td>1997</td>
<td>- + - - - + + + + +</td>
<td>+ + - + - - - + - +</td>
<td>2</td>
<td>4</td>
<td>46%</td>
</tr>
<tr>
<td>Hirata</td>
<td>1997</td>
<td>+ + + - - + + - + -</td>
<td>+ + - + + + + - + +</td>
<td>4</td>
<td>4</td>
<td>62%</td>
</tr>
<tr>
<td>Isenberg</td>
<td>1998</td>
<td>- + - - - - - + - -</td>
<td>+ + - + - - - - - +</td>
<td>3</td>
<td>4</td>
<td>54%</td>
</tr>
<tr>
<td>Bowrey</td>
<td>1999</td>
<td>- - + - - - + + + +</td>
<td>+ + - + - - - - - +</td>
<td>2</td>
<td>5</td>
<td>54%</td>
</tr>
<tr>
<td>Laterza</td>
<td>1999</td>
<td>- - + - - - + + + +</td>
<td>+ + - + - - - - - +</td>
<td>3</td>
<td>4</td>
<td>54%</td>
</tr>
<tr>
<td>Zuccaro</td>
<td>1999</td>
<td>- - + + + - + + + +</td>
<td>+ + - + - - - + + +</td>
<td>4</td>
<td>5</td>
<td>69%</td>
</tr>
<tr>
<td>Beseth</td>
<td>2000</td>
<td>- - + + - + - + + -</td>
<td>- - - + + + + + + +</td>
<td>1</td>
<td>3</td>
<td>31%</td>
</tr>
<tr>
<td>Chak</td>
<td>2000</td>
<td>- + - - + - + + + +</td>
<td>+ + - + - - - - + +</td>
<td>3</td>
<td>5</td>
<td>62%</td>
</tr>
<tr>
<td>Willis</td>
<td>2002</td>
<td>+ + + + + - + + + +</td>
<td>+ + + + + + + + + +</td>
<td>6</td>
<td>5</td>
<td>85%</td>
</tr>
<tr>
<td>Kroep</td>
<td>2003</td>
<td>+ + - + + - + + + +</td>
<td>+ + - + + + + + + +</td>
<td>5</td>
<td>5</td>
<td>77%</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>54%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>PET Internal Validity</th>
<th>PET External Validity</th>
<th>Total IV</th>
<th>Total EV</th>
<th>% of Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arslan</td>
<td>2002</td>
<td>+ + - - + + + + - + +</td>
<td>+ + + + + + + + + + +</td>
<td>4</td>
<td>4</td>
<td>62%</td>
</tr>
<tr>
<td>Kato</td>
<td>2002</td>
<td>+ + + - - + + + + +</td>
<td>+ + + + + + + + + +</td>
<td>4</td>
<td>3</td>
<td>54%</td>
</tr>
<tr>
<td>Downey</td>
<td>2003</td>
<td>- + + - + + + + + + +</td>
<td>+ + - - - + + + + + +</td>
<td>5</td>
<td>2</td>
<td>54%</td>
</tr>
<tr>
<td>Brucher</td>
<td>2001</td>
<td>+ + + + - + + + + - +</td>
<td>+ + - + - + + + + + +</td>
<td>6</td>
<td>5</td>
<td>85%</td>
</tr>
<tr>
<td>Weber</td>
<td>2001</td>
<td>+ + + + - - + + + + +</td>
<td>+ + - - + + + + + + +</td>
<td>6</td>
<td>5</td>
<td>85%</td>
</tr>
<tr>
<td>Flamen</td>
<td>2002</td>
<td>+ + + - - + + + + + +</td>
<td>+ + - + + + + + + + +</td>
<td>6</td>
<td>6</td>
<td>92%</td>
</tr>
<tr>
<td>Kroep</td>
<td>2003</td>
<td>+ + + + + + + + + + +</td>
<td>+ + - + + + + + + + +</td>
<td>7</td>
<td>6</td>
<td>100%</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>85%</td>
</tr>
</tbody>
</table>

IV1-IV7: 7 criteria for Internal Validity; EV1-EV6: 6 criteria for External Validity; Total IV: total score of Internal Validity; Total EV: total score of External Validity; The + and – symbols indicate the presence or absence of the criteria for internal or external validity.
RESULTS

Literature search
With our initial search, we identified 58 studies about CT, 26 studies about EUS, and 26 studies about FDG-PET. After reviewing the title and the abstract of these studies, some studies were excluded: 53 studies with CT, 12 studies with EUS, and 15 studies with FDG-PET. These studies included narrative reviews, case reports, studies in which researchers reported the use of these modalities for initial staging, and studies in which carcinomas other than esophageal cancer were included.

Five studies were potentially relevant to the value of CT in the monitoring of a response. Of these remaining five studies, the study by Helmberger et al\(^\text{18}\) was excluded because of the absence of a valid reference standard. In that study, CT during neoadjuvant therapy was compared with endoscopy, although the value of endoscopy in the monitoring of the response to therapy is still debatable. Thus, only four studies with CT met the inclusion criteria\(^{19-22}\).

In the case of EUS, 14 studies were potentially relevant to the value of therapeutic response assessment. The study by Nousbaum et al\(^{23}\) was excluded because of the absence of a valid reference standard; the authors used recurrence instead of histologic findings as the reference standard. Thus, 13 studies with EUS met the inclusion criteria\(^{21,24-35}\).

For studies with FDG-PET, 11 were potentially relevant to the value of therapeutic response assessment. Of these 11 studies, four were excluded after a review of the full article. One study was excluded because of the absence of a baseline FDG-PET image.\(^{36}\) Couper et al compared FDG-PET during neoadjuvant therapy with CT and changes in weight and dysphagia scores.\(^{37}\) Since the value of CT in monitoring response is still under discussion, it is not suitable to be used as a reference standard. Another two studies were excluded because one comprised carcinomas other than esophageal cancer and one included fewer than 10 patients with esophageal cancer.\(^{38,39}\) Thus, seven studies met the inclusion criteria.\(^{21,40-45}\)

The characteristics of the included studies are shown in Table 2.

Methodological quality assessment
Thirteen methodological quality items were assessed for each of the 24 studies (Table 3). There was disagreement in 20 out of 312 items that were assigned a score. Disagreement was most frequent for internal validity criterion 4 and internal validity criterion 5, mainly because of reading errors and/or differences in interpretation. The percentage of maximum score for the combined internal and external validity ranged from 15% to 100%, with a median of 62% for studies with CT, 54% for those with EUS, and 85% for those with FDG-PET.

Most studies were assigned a positive score for avoidance of verification bias (internal validity criterion 5 with positive score: 22 of 24 studies = 92%), inclusion of right spectrum of disease (external validity criterion 1 with positive score: 22 of 24 studies = 92%), and provision of sufficient demographic information (external validity criterion 2 with positive score: 21 of 24 studies = 88%). For most studies, performance was poor in regard to avoidance of selection bias because a consecutive series of eligible patients was not included (external validity criterion 5 with positive score: seven of 24 studies = 29%).

Quality of studies with FDG-PET was high for definition of the reference test (internal validity criterion 1: six of seven = 86% versus three of four = 75% for studies with CT and only...
four of 13 = 31% for studies with EUS) and definition of the index test (internal validity criterion 2: seven of seven = 100% versus only six of 13 = 46% for studies with EUS). The latter item was also present in all studies with CT.

**Studies with CT**

Walker et al evaluated 38 patients with esophageal cancer by using CT and a barium swallow study before and after completion of preoperative chemotherapy. Chemotherapy consisted of a combination of fluorouracil, doxorubicin hydrochloride, (Adriamycin), and mitomycin (n = 15 patients), also known as FAM; a combination of mitomycin, ifosfamide, and cisplatin (n = 19 patients), also known as MIC; or a combination of cyclophosphamide, doxorubicin hydrochloride, vincristine sulfate, methotrexate, and etoposide (n = 4), also known as CAPOMET. Radiographic response was assessed according to the criteria of Miller et al. Pathologic response was defined as follows: complete microscopic response (no microscopic evidence of residual tumor); complete macroscopic response (no visible evidence of residual tumor); partial response (unequivocal signs of healing); no response (no substantial signs of healing). A response to therapy, either partial or complete, was found in 94% (n = 36) of the patients, and 48% (n = 18) of all patients showed a response at CT.

Griffith et al assessed the therapeutic response by using spiral CT after completion of two courses of fluorouracil and cisplatin in 45 patients with squamous cell carcinoma. Response to therapy was defined as a volume reduction of more than 50%. The pathologic response was assessed according to the criteria of Mandard et al. According to these criteria, 24% (n = 11) of patients were responders. Griffith et al did not find a correlation between tumor volume reduction at serial CT and quantitative pathologic tumor assessment, nor did they find a correlation between tumor volume reduction and survival.

Jones et al assessed the value of CT for tumor response assessment in 50 patients with esophageal cancer who were treated with fluorouracil and cisplatin, whereas in 12 patients, paclitaxel was added to the regimen. Radiographic response was determined by using the response criteria for solid tumors of the Eastern Cooperative Oncology Group. Pathologic response was defined as follows: no tumor (42% for responders) and tumor present (58% for nonresponders). CT had a sensitivity of 33% and a specificity of 66% in accurate evaluation of the pathologic response of the tumor to chemoradiation therapy.

Kroep et al evaluated 13 patients with esophageal cancer who were treated with neoadjuvant cisplatin and gemcitabine plus granulocyte macrophage colony-stimulating growth factor. CT assessment of the therapeutic response was determined, according to World Health Organization criteria. A pathologic response was observed in 36% of the patients, as assessed according to criteria of Mandard et al. Both early (after two cycles of neoadjuvant therapy) and late (after completion of neoadjuvant therapy) response evaluation showed a specificity of 50% and a sensitivity of 71%.

**Studies with EUS**

In nine studies with EUS that were included, restaging, which was indicated by a change in tumor stage, was used as a parameter for assessment of neoadjuvant therapeutic response. In three EUS studies, change in volume measurements of the maximum tumor cross-sectional dimensions was used as a parameter for assessment. Researchers in one study evaluated both restaging and volume measurements. These are different parameters and, thus,
evaluation of the results was interpreted separately.

Restaging
In seven of nine studies, a test definition was not described in regard to restaging with EUS after neoadjuvant therapy.24,25,27,30-33,35 Before and after therapy, staging with the TNM classification was described properly, but the definition of responders or nonresponders was not described. Therefore, the value of EUS for the monitoring of a response in these studies could not be determined. Accuracy in regard to tumor staging after completion of therapy, however, can be determined and ranged from 27% to 82%, with a median of only 48%.

Giovannini et al described a test definition.28 Thirty-two patients were included in that study. Complete EUS was not performed in six of the 32 patients before therapy because of esophageal stenosis. After therapy, complete EUS was performed in all patients. Thus, findings at EUS in 26 patients with T3 and T4 esophageal carcinomas who were treated with chemoradiotherapy were correlated to the histologic findings in the resected specimens. Chemoradiotherapy consisted of two courses of fluorouracil and cisplatin and a total of 30 Gy of radiation therapy. EUS criteria for prediction of a response were as follows: T0, complete response; T1, microscopic evidence of tumor residue (major response); and T2-T4, no response, or minor response (depending on the primary stage prior to chemoradiotherapy). If the histologic findings in the resected specimen indicated no tumor (complete responder) or pT1 tumor (partial responder), patients were defined as responders.

Kroep et al evaluated 13 patients with esophageal cancer who were treated with neoadjuvant cisplatin and gemcitabine plus granulocyte macrophage colony-stimulating growth factor.21 At EUS, the therapeutic response assessment was defined as complete if there was no visible tumor left, as partial after downstaging with the TNM criteria, as progressive in case of an increase in the TNM stage, or as stable when no change in TNM stage was determined. Pathologic response was assessed according to criteria of Mandard et al.47 The early (after 2 cycles) and late (after completed induction therapy) responses showed a specificity of 100% and a sensitivity of 100% in 12 patients who could be evaluated. Evaluation was not feasible in one patient due to development of liver metastases, which precluded surgical resection. In one patient, EUS after treatment was not feasible because of the impossibility of esophageal passage caused by the tumor. This patient, however, was classified as a nonresponder and not excluded from the analysis.

Volume measurement
Isenberg et al conducted a study in 31 patients to determine whether the estimation of tumor size at EUS could be used to assess the response to preoperative chemoradiation (cisplatin and carboplatin with fluorouracil and concurrent mediastinal radiation of at least 30 Gy).31 An EUS response was defined as a 50% reduction in maximal cross-sectional area of the tumor. A shortcoming of this study is the absence of a definition of pathologic responders and nonresponders, the lack of which makes it impossible to reproduce the data.

Chak et al measured maximal cross-sectional area in 50 patients before and after neoadjuvant therapy (cisplatin or carboplatin and fluorouracil concurrently administered with a total of 50 Gy radiation therapy).26 Patients with more than a 50% reduction were classified as EUS responders. The authors compared survival and not pathologic findings, with patients classified as responders and those classified as nonresponders with EUS. Survival time
in the group of responders was significantly longer than it was in the group of nonresponders (median survival of 17.6 months versus 14.5 months, \( P < 0.05 \)).

Hirata et al evaluated pre- and posttherapeutic EUS in 34 patients who underwent various preoperative treatments consisting of radiation therapy (to a total of 30 Gy), chemotherapy (bleomycin or cisplatin), and hyperthermia.\(^{29}\) In 17 patients, the tumor could not be passed before treatment; in these patients, the maximum area was measured only at the level of the most cranial portion. Response at EUS was defined as a 15% to 30% reduction and more than a 30% reduction in maximal cross-sectional area. Pathologic data were available of 27 patients who underwent surgery. Histopathologic response was graded according to the number of viable cells in the entire lesion. Therapy was considered ineffective when viable cells occupied more than two-thirds of the entire tumor. Hirata et al found a correlation between percentage of reduction in maximal cross-sectional area and histologic evidence of effectiveness of neoadjuvant therapy. There was a significant difference (\( P = 0.05 \)) in overall survival rates among the three groups, according to a reduction in the area of the tumor: less than 15%, 15% to 30%, and more than 30%.

Willis et al correlated responses as measured by EUS in 41 patients with a defined pathologic tumor regression grade as proposed by Mandard et al.\(^{34,47}\) Patients received cisplatin or carboplatin and fluorouracil concurrently with a total of 50 Gy of radiation therapy. A positive response at EUS was defined as more than a 50% reduction in maximal cross-sectional area. Willis et al found a strong correlation between the EUS therapeutic response and chemoradiotherapy-induced pathologic tumor regression.

Studies with FDG-PET
Arslan et al evaluated the value of the use of FDG-PET in 24 patients who underwent chemoradiation therapy consisting of administration of cisplatin/fluorouracil, cisplatin/taxotore, or carboplatin/fluorouracil and concurrent radiation therapy with 50.4 Gy.\(^{40}\) The response of the primary tumor was visually assessed and analyzed semiquantitatively with SUV before and after treatment. Patients were classified into two groups according to histopathologic findings at surgery: responding patients without gross disease and nonresponding patients with gross disease. Differences in quantitative FDG-PET data and between responders and nonresponders were calculated as a mean value for both groups and not on a per-patient basis. Therefore, accuracy could not be assessed.

Kato et al described 10 patients who received nedaplatin and fluorouracil in combination with radiation therapy of 40 Gy.\(^{44}\) Histologic response, according to the Japanese Society of Esophageal Diseases, was correlated with the SUV after treatment and the FDG uptake length decrease; however, it was not correlated with the rate of SUV decrease.\(^{48}\)

Downey et al assessed the prognostic value of FDG-PET in 17 patients with esophageal carcinoma who were treated with chemotherapy (paclitaxel/cisplatin), radiotherapy (50.4 Gy), and surgery.\(^{42}\) At a threshold of a 60% decrease in FDG uptake, they found a significant difference (\( P = 0.055 \)) in 2-year disease-free survival between patients who had less decrease in SUV (38% survival) and patients who had a greater decrease in SUV (67% survival).

In a study of Brucher et al, 27 patients were included who underwent external-beam radiation therapy and treatment with fluorouracil as a continuous infusion.\(^{41}\) Pathologic response was assessed according to the criteria of Mandard et al.\(^{47}\) At a threshold level of
52% decrease in FDG uptake compared with the baseline value, sensitivity for prediction of a response was 100%, with a corresponding specificity of 55%. The positive and negative predictive values were 72% and 100%, respectively. Median survival was $22.5 \pm 2.4$ in responders evaluated with FDG-PET versus only $6.7 \pm 5$ in nonresponders evaluated with FDG-PET ($P < 0.001$).

Weber et al reported a series of 40 patients who received chemotherapy of 72 days duration. After 14 days of therapy, mean reduction of tumor FDG uptake, according to criteria of Mandard et al, was significantly different between responders (54% ± 17%) and nonresponders (15% ± 21%). An optimal differentiation was achieved by using a cutoff value of 35%. The 2-year survival rate was 60% in responders evaluated with FDG-PET versus 37% in nonresponders evaluated with FDG-PET ($P < 0.04$).

Flamen et al included 36 patients in their study with T4 esophageal cancer, who received concomitant external-beam radiation therapy plus cisplatin and fluorouracil. Patients were classified as major responders evaluated with FDG-PET when a strong reduction in FDG uptake was demonstrated at posttherapy PET. A patient was classified as a major responder to therapy when the histologic findings of the primary tumor in the resection specimen indicated a classification of pT0-pT2 or pT3 (if the residual tumor consisted of small foci of viable tumor on a background of extensive histopathologic response) and there was no sign of any tumoral viability beyond the primary tumor site. After patients completed chemoradiotherapy, the sensitivity of serial FDG-PET for a major response was 71% (10 of 14 patients), and specificity was 82% (18 of 22 patients). Major responders evaluated with FDG-PET had a median survival of 16.3 months, and nonmajor responders, 6.4 months ($P=0.005$).

Kroep et al evaluated 13 patients with esophageal cancer who were treated with neoadjuvant cisplatin and gemcitabine plus granulocyte macrophage colony-stimulating growth factor. The pathologic response was assessed according to criteria of Mandard et al. Evaluation of early (after two cycles) and late (after completed induction therapy) response yielded a specificity of 86% and 100%, respectively, and a sensitivity of 100% at both intervals, with cutoff values of 40% for early responses and 60% for late responses.

Comparison of diagnostic accuracy

Three studies with CT, four with EUS, and four with FDG-PET provided enough data to permit calculation of sensitivity and specificity (Figure 1). The sensitivity of CT, EUS and FDG-PET ranged from 33% to 55%, 50% to 100%, and 71% to 100%, respectively. The specificity ranged from 50% to 71%, 36% to 100%, and 55% to 100%, respectively. Summary ROC curves for the three modalities are given in Figure 2. The maximum joint sensitivity and specificity (Q-point) values for CT, EUS, and FDG-PET were 54% (95% confidence interval (CI): 31% to 77%), 86% (95% CI: 80% to 93%), and 85% (95% CI: 77% to 93%), respectively. Overall accuracy of CT was significantly lower than was the accuracy of EUS ($P < 0.003$) and of FDG-PET ($P < 0.006$). Overall accuracy of EUS was similar to that of FDG-PET ($P = 0.84$). The estimated specificity values of CT, EUS and FDG-PET (values that corresponded to a desired level of sensitivity of 90%) were 13%, 78%, and 78%, respectively. Evaluation of neoadjuvant therapy in the analyzed studies was not feasible by using PET in 1 (< 1%) of 116 patients, and it was not feasible by using EUS in seven (6%) of 120 patients. EUS was suboptimal in 17 (14%) of 120 patients. In contrast, CT was always feasible.
FIGURE 1

Forest plots of sensitivities (A) and specificities (B) for all 11 studies. Plots show variation in estimates between studies and modalities. Confidence intervals are large in the majority of studies because of small sample sizes. (EUS = endoscopic ultrasonography, TP = true positives; FN = false negatives; TN = true negatives; FP = false positives)
DISCUSSION

To determine the best diagnostic tool for the evaluation of the response to neoadjuvant therapy response in patients with esophageal cancer, we systematically reviewed the literature in regard to the role of CT, EUS, and FDG-PET. The included studies showed limited methodological quality. In our review, researchers in all studies with CT reported a low accuracy for the assessment of response in esophageal cancer, and this was significantly lower than that of FDG-PET and EUS. In the studies with FDG-PET and EUS, accuracy was good end equivalent for both modalities.

CT generally is considered the state-of-the-art diagnostic modality for monitoring nonsurgical therapy in solid tumors. For therapeutic response assessment in esophageal cancer, however, accuracy was low with CT. This is probably related to the difficulty in the differentiation between viable tumor and reactive changes, including edema and scar tissue, at CT. Another reason may be the knowledge that CT is of limited value for initial tumor staging, because of poor differentiation between tumors with staged to T3. It is far less accurate than is EUS, which is presently considered the standard for staging of the primary tumor. Although only a limited number of studies with far from optimal methods were available, we conclude that single-detector row CT is a poor diagnostic tool for the determination of the pathologic tumor response after chemoradiotherapy in patients with
Review of Response Assessment in Esophageal Cancer

esophageal cancer.

The articles with CT included in this review concerned studies in which thick collimation was used. With the introduction of multi-section scanners, the use of thinner collimation has become possible. Thinner collimation can be expected to lead to improved delineation of the tumor and, hence, improved 3D measurements and more accurate and reproducible measurements of tumor volume. This may improve the results of CT in the near future for monitoring of the response with respect to tumor volume.

EUS has been developed over the past decade and has been shown to be highly accurate for the initial staging of the primary tumor. In contrast to staging of untreated esophageal cancer, the value of restaging with EUS after neoadjuvant therapy appears to be limited by the difficulty in the differentiation of residual carcinoma from inflammation and fibrosis.\textsuperscript{30,35} This is supported by the finding that restaging with EUS after therapy had a poor accuracy, with pathological staging as the reference standard; accuracy ranged from 27% to 82%. Nevertheless, the accuracy of EUS for therapeutic response assessment in our review was significantly higher than the accuracy of CT and was comparable to the accuracy of FDG-PET. These results should be interpreted with caution, because they were based on four studies with poor to moderate methodological quality.

Factors that may limit EUS for assessment of a nonsurgical response are postradiation esophagitis, luminal stenosis, and compression of the tumor caused by the endoscope.\textsuperscript{30,49} In the four studies with EUS that were included for ROC analysis in this review, EUS was not feasible in 6%\textsuperscript{21,28} and was suboptimal in 14% of patients.\textsuperscript{29} In addition, to be useful for therapeutic response assessment, performance of repeated measurement with high reproducibility is of crucial importance, and this can only be achieved if well-trained and experienced operators are available.

Overall accuracy of FDG-PET was similar to that of EUS. The maximum joint sensitivity and specificity was 85% and 86%, respectively. Metabolic imaging with FDG-PET enables discrimination of viable tumor from necrotic scar tissue in patients with esophageal cancer after neoadjuvant therapy. Nonspecific glucose uptake by tissue with inflammation caused by chemotherapy and/or radiation therapy may be a limitation for FDG-PET in therapeutic response assessment. Although the suboptimal sensitivity by false-negative results for response assessment in some studies\textsuperscript{41,43} could be explained by this mechanism, this finding is rare and not observed in other studies.\textsuperscript{21,45} Most likely, the contribution of FDG uptake caused by inflammation to total FDG uptake is low, and, moreover, the threshold for FDG decrease to define responders can be adjusted to select responders almost exclusively (i.e. sensitivity of 100% with 100% negative predictive values only as performed by Weber et al and Kroep et al).\textsuperscript{50}

A wide range of cutoff values, from 30% to 80% and expressed as SUV, for reduction in FDG uptake has been reported for use in the discrimination between responders and nonresponders. This range of cutoff values suggests a lack of standardization, but it also might be explained by factors such as spectrum of disease, differences in therapy (e.g. severity in tumor kill, and timing of evaluation of the therapy). Kroep et al used an optimal cutoff value of 40% after two cycles and of 60% after completion (four to six cycles) of chemotherapy. Still, an advantage of FDG-PET is the high reproducibility, with a change of greater than 20% considered to reflect a true biological effect.\textsuperscript{50-53}

Furthermore, various analytical models (SUV with corrections for body surface...
area, simplified kinetic model, nonlinear regression analysis, etc), are available and have been tested to optimize therapeutic response monitoring by using FDG-PET. These models have been described in detail by Hoekstra et al and have been described particularly for esophageal cancer by Kroep et al. Because assessment of SUV combines accuracy with simplicity, this model is used in most studies. Further research and validation, focusing on the cutoff values with the different analytical models for optimal discrimination between responders and nonresponders, has to be done.

For this review, we used histopathologic findings as a valid reference standard. The ultimate aim, however, was to identify patients with a poor clinical outcome early in the course of neoadjuvant therapy; in only seven studies was survival reported, and these studies included one with CT, two with EUS, and four FDG-PET. In the study with CT, outcome was not predicted, whereas in both studies with EUS, a slightly but significantly higher survival was found in responders versus nonresponders. In all four studies with FDG-PET, striking differences with significantly longer survival in metabolically responding patients were found.

Response to neoadjuvant treatment was evaluated after completion of this nonsurgical treatment in most studies. These findings may hold prognostic value as mentioned previously, but they will not alter the duration or intensity of neoadjuvant treatment. Only if evaluation is performed early in the course of neoadjuvant treatment and the test may accurately discriminate responders from nonresponders can the test results be used to aid in the decision about whether this toxic therapy should be continued, as in responders, or stopped, as in nonresponders. Researchers in only two studies investigated the role of FDG-PET in distinguishing responders from nonresponders early during the course of chemoradiotherapy. Kroep et al and Weber et al showed encouraging results, with high negative predictive values for the early response to neoadjuvant treatment in patients with esophageal cancer (95% and 100%, respectively).

Early prediction of tumor response might be improved by taking other factors into account. For esophageal cancer, several histologic markers, such as the tumor suppressor gene p53, the proliferative marker Ki-67, and the epidermal growth factor receptor have been evaluated for the prediction of the therapeutic response even prior to neoadjuvant therapy. Neither a single marker nor a combination of markers, however, can currently be used to predict the response with sufficient accuracy. In the future, gene expression profiling may identify markers that can be used in combination with imaging results for prediction of the response to neoadjuvant therapy.

In addition, there are some general limitations to this review, which should be kept in mind when one interprets the summary ROC curves. The number of studies and patients included in our review was small. Only 11 studies for the three modalities combined were eligible for ROC analysis, which provided a total of 318 patients.

In none of the studies was a head-to-head comparison used to test directly for a difference in accuracy between modalities within the same group of patients. Although Kroep et al evaluated all three modalities in the same patient population, they did not directly compare accuracy, possibly because of the small number of patients who could be evaluated and who underwent examinations with all three modalities.

Some potential sources of heterogeneity need to be addressed. First, neoadjuvant therapy schemes differed substantially among the studies, with a major difference in the
use of chemotherapy alone in four studies versus the combination of chemotherapy and radiation therapy in 7 studies. Second, differences in the spectrum of disease can be an important source of heterogeneity. We restricted our review to studies that included patients who were eligible for curative surgery and, therefore, were candidates for neoadjuvant treatment. Nevertheless, tumor stage may vary from small T1 tumors to larger T3 tumors. Measurements of change in tumor volume or of FDG uptake may be less accurate in small tumors than they are in large tumors because of the larger influence of systematic errors and lower statistics. Moreover, small tumors may be missed at CT or FDG-PET at initial staging. Subanalyses in future larger studies are necessary to document the influence of initial tumor stage on therapeutic response assessment. Third, it is important to know whether researchers defined their cutoff value prior to the study or after they obtained the results. In the latter case, there is an increased likelihood that the authors selected the cutoff value to maximize a particular test characteristic, which reduces the likelihood that investigators in another study will be able to replicate the findings. Different methods or thresholds to define a positive test result can lead to a special source of variation. Response criteria were defined differently within each modality (Table 2). Two different parameters were used for EUS (i.e. restaging and maximal cross-sectional area). In all studies with CT, volume measurements were used, but the thresholds differed slightly. In all studies with FDG-PET, except that of Flamen et al, the SUV was used. Flamen et al used the delta tumor-to-liver uptake ration, which is comparable with the SUV, albeit with correction for liver activity. These differences, however, are small, and the summary ROC approach was developed to incorporate differences in accuracy that arise because of a change in threshold between studies for defining a test as positive.

In conclusion, single-section CT for assessment of the response to neoadjuvant therapy in esophageal cancer is inaccurate and, therefore, is not recommended. EUS and FDG-PET have equivalent accuracy. EUS can be used to identify patients in whom a pathologic response was achieved, but is not always feasible during or shortly after chemoradiation and, therefore, is not routinely used for therapeutic response assessment. FDG-PET, which is used to measure alterations in tissue metabolism, seems to be a promising noninvasive tool for the assessment of the response to neoadjuvant therapy in patients with esophageal cancer. Larger studies with sufficient power that focus on the prediction of the tumor response early in the course of neoadjuvant therapy and that include direct comparisons of different modalities are needed.
REFERENCES


