Background and aims: After an initial period of excellent results with newly introduced imaging procedures, the accuracy of most imaging methods declines in later publications. This effect may be due to various methodological factors involved in the research. Using the example of endoscopic ultrasound (EUS), this study aimed to elucidate one of the factors possibly concerned—namely, the extent to which the examiners are adequately blinded.

Methods: Well documented videotapes of EUS examinations of 101 patients with resected tumours of the oesophagus (n=32), stomach (n=33), or pancreas (n=36) were evaluated in three different ways: firstly, retrospective analysis under routine clinical conditions; secondly, evaluation of EUS videotapes in a strictly blinded fashion; and thirdly, evaluation of the same videotapes but with additional information from the video endoscopic appearance (oesophageal/gastric cancer) or from computed tomography results (pancreatic cancer). Histopathological T staging was used as the reference method.

Results: The accuracy of EUS in T staging was 73% under routine conditions. This value fell significantly to 53% for the blinded evaluation but increased again to 62% for the unblinded evaluation. The sensitivity of staging T1/T2 tumours was 72% (routine EUS), 59% (blinded EUS), and 70% (unblinded EUS). The respective values for advanced tumours were 85%, 74%, and 72%.

Conclusions: The accuracy of EUS for T staging in clinical practice appears to be lower than has previously been reported. In addition, blinded analysis produced significantly poorer results, which improved when another test was added. It may be speculated that better results with routine EUS obtained in a clinical setting are due to additional sources of information.
(2) Blinded analysis. The videotapes recorded at the initial examination were re-evaluated by one of the main investigators (TR). Patients were mixed, and their names were concealed. This analysis was performed a minimum of 18 months (range 18–24 months) after the initial routine examination. At this time, the only information available to the examiner was the anatomical site of the tumour. Other data, such as the patient's name, age, sex, and clinical findings, were not revealed.

(3) Unblinded analysis. A minimum of a further 18 months (range 18–24 months) after the blinded re-evaluation, the videotapes recorded at the initial examination were evaluated once again. However, this time the investigator was not blinded, and was allowed to review the corresponding endosonographic tapes before the EUS assessment (for oesophago gastric cancers) or to read the CT reports (for pancreatic cancers). Other clinical data (including patient names), and any imaging information other than the endoscopy or CT data, were still not made available.

EUS examinations were done in a standardised way by all examiners by scanning the tumour extent after passage to the distal tumour free part of the organ and then slowly withdrawing the instrument along the tumour axis; similarly, established and previously published criteria for EUS staging were used. Briefly, oesophago gastric tumours mostly appeared as echo poor localised or circumferential wall thickening, with progressive loss of the underlying wall layer structure corresponding to the tumour stage. Stage T1 was thus assumed on EUS if the lesion was limited to the mucosa and submucosa (first three layers). Stage T2 was assumed if the tumour infiltrated the entire wall, leading to complete or nearly complete loss of the layer structure, but with a smooth outer margin and/or small parts of the muscularis layer (fourth layer) being recognisable. In contrast, the echo poor wall thickening leading to an endosonographic diagnosis of stage T3 showed irregular outer margins, with clear infiltration of the adventitia and surrounding fat tissue. In stage T4, infiltration of other organs had to be visualised. For pancreatic cancer, the old TNM system (1987) was used: stage T1 was assumed on EUS if the tumour was limited to the pancreas, and stage T2 if it had spread beyond the pancreas but was not infiltrating major parapancreatic vessels, as was the case when stage T3 was diagnosed. Criteria for vascular infiltration were total or partial vascular obstruction, with or without collateral vessels, visualisation of tumour in the vessel, or a grossly irregular tumour-vessel interrelationship.

The results obtained using all three assessment methods were compared with the histopathological findings for the resected tumours, which served as the gold standard. Only resected cases and those pancreatic cancers with thorough surgical exploration were included in order to have a reliable and uniform gold standard.

To assess interobserver variability as a potentially confounding factor, oesophago gastric tumour assessments for the blinded and unblinded analyses were repeated by a second independent examiner (HJD) under the same conditions. Both evaluators (TR and HJD) had experience of more than 500 oesophago gastric and more than 250 pancreatic EUS examinations at the time of the study.

Accuracy rates were calculated for all three methods. In addition, the sensitivity of EUS for detecting less invasive (T1/2) or advanced tumours (T3/T4) was determined. Differences between the respective groups were tested for statistical significance using the McNemar test for paired samples. A p value <0.05 was considered significant.

RESULTS

After screening of the relevant videotape sequences from the study period, a total of 101 patients were identified who met the entry criteria. The results for the three assessment methods showed an overall T staging accuracy of 73.3% (74 of 101) for routine EUS examinations which was significantly better than the rate of 52.5% (53 of 101) for the blinded evaluation of the videotapes and 62.4% (63 of 101) for the unblinded re-evaluation of the videotapes (McNemar test: p < 0.001 and p = 0.043, respectively) (fig 1). However, the “unblinded” evaluation involved adding information from only one other imaging test (endoscopy or CT) but no other clinical or imaging data. The information available in clinical routine practice may therefore still contain substantially more data than provided in this “unblinded” analysis.

When the data for localisation of tumours were stratified, it was found that blinded evaluation of the EUS videotapes was associated with a reduced accuracy for all three tumour sites, reaching statistical power for the oesophagus and stomach. Although the accuracy rates for pancreatic and oesophago gastric cancers improved significantly after the investigator was “unblinded” before re-evaluating the videotapes, no such effect was seen for gastric tumours (table 1). This low level of accuracy in gastric cancer even when the examiner was not blinded involved overstaging of the tumours in most cases (14 of 20). Overstaging was particularly frequent with pT2 tumours (nine of 12) and those located in the proximal stomach (six of eight).

The accuracy rates for the various T stages (T1–T3), as evaluated by the three methods, are shown in table 2. Only one

**Table 1** Overall accuracy of T staging, stratified according to the location of the cancer

<table>
<thead>
<tr>
<th>Cancer Location</th>
<th>Routine analysis</th>
<th>Blinded</th>
<th>Unblinded</th>
</tr>
</thead>
<tbody>
<tr>
<td>All patients (n=101)</td>
<td>73.3%</td>
<td>52.5%**</td>
<td>64.2%*</td>
</tr>
<tr>
<td>Oesophageal cancer (n=32)</td>
<td>81.3%</td>
<td>50.0%**</td>
<td>71.9%</td>
</tr>
<tr>
<td>Gastric cancer (n=33)</td>
<td>66.7%</td>
<td>45.5%*</td>
<td>39.4%**</td>
</tr>
<tr>
<td>Pancreatic cancer (n=36)</td>
<td>72.2%</td>
<td>61.1%</td>
<td>75.0%</td>
</tr>
</tbody>
</table>

*p < 0.05, **p < 0.01 compared with “routine analysis” (McNemar test).
patient with a gastric cancer had a tumour in stage pT4 (he underwent palliative resection) which was correctly staged in the routine EUS examination and in the blinded evaluation of the tapes, and was understaged as a T3 tumour after unblinded re-evaluation.

When the sensitivity of EUS for detecting tumours limited to the wall (T1/T2) was compared with its sensitivity for detecting advanced tumours (T3/T4), it was found that all three methods had better sensitivity rates in staging advanced lesions. However, both blinded and unblinded re-evaluation of the videotapes led to lower accuracy rates compared with those obtained in the routine analysis. This difference proved to be significant for both groups in relation to stage T3/T4 tumours. For non-invasive stage T1/T2 tumours, only the blinded analysis resulted in a significantly lower level of accuracy (table 3).

In patients with oesophageal tumours, excluding those who had a non-traversable tumour stenosis (n=13), did not lead to improved accuracy with EUS (routine analysis 78.9% v 81.3%; blinded analysis 42.1% v 50.0%; unblinded analysis 68.4% v 71.9%).

A second examiner (HJD) was involved in reviewing the videotapes for the staging of oesophagogastric tumours, and his findings were compared with those of the main investigator (TR); the same effects were observed. The respective kappa values for testing interobserver variability between the two examiners were 0.264 (blinded) and 0.528 (unblinded).

DISCUSSION
The main aim of the present paper was not to present accuracy or outcome data on endosonographic cancer staging but to further analyse some of the methodological factors involved in scientific publishing concerning the accuracy of imaging tests. We speculated that one of those factors may be additional information being fed into the seemingly “pure” interpretation of the test results. This could be due to the simple fact that the investigators are familiar with the patients concerned and their clinical features, indicating advanced disease; the results of blood tests, previous endoscopic findings, or results previously obtained from other tests are usually known before an examination is begun. While this total view of the patient and his or her findings, taking into account every aspect of the disease, is an excellent approach for clinical case management, it is problematic during scientific evaluation of a specific procedure or subset of procedures when the aim should be to have a minimum of confounding factors. Particularly when the results of imaging tests are contradictory in clinical practice, the “real” accuracy of each procedure needs to be known independently of other procedures.

There is ample evidence of the phenomenon of initial good results which are not maintained in later publications, both in the area of EUS—the topic of the present study—and also with other imaging procedures. This was recently shown for EUS in pancreatic cancer staging.5–7 In addition, comparisons

### Table 2
Accuracy of T staging, stratified according to pT stages T1–T3

<table>
<thead>
<tr>
<th></th>
<th>All patients (n=100)</th>
<th>pT1 (n=10)</th>
<th>pT2 (n=44)</th>
<th>pT3 (n=46)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine analysis</td>
<td>73.0%</td>
<td>60.0%</td>
<td>63.6%</td>
<td>84.8%</td>
</tr>
<tr>
<td>Blinded</td>
<td>52.0%**</td>
<td>60.0%</td>
<td>43.2%*</td>
<td>58.7%*</td>
</tr>
<tr>
<td>Unblinded</td>
<td>63.0%*</td>
<td>40.0%</td>
<td>61.4%</td>
<td>69.6%*</td>
</tr>
</tbody>
</table>

*p<0.05, **p<0.01 compared with “routine analysis” (McNemar test).

### Table 3
Sensitivity of endoscopic ultrasonography for detecting pT1/T2 cancers versus pT3/T4 cancers

<table>
<thead>
<tr>
<th></th>
<th>pT1/T2 (n=54)</th>
<th>pT3/T4 (n=47)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine analysis</td>
<td>72.2%</td>
<td>85.1%</td>
</tr>
<tr>
<td>Blinded</td>
<td>59.3%*</td>
<td>74.5%†††</td>
</tr>
<tr>
<td>Unblinded</td>
<td>70.4%</td>
<td>72.3%†††</td>
</tr>
</tbody>
</table>

*p<0.05 compared with “routine analysis” and “unblinded” (McNemar test); †††p<0.001 compared with “routine analysis” (McNemar test).

Figure 2  An oesophageal carcinoma in histopathological stage T2 but with endosonographic visualisation of transmural tumour growth, falsely suggesting stage T3 on endoscopic ultrasonography.

Figure 3  An oesophageal carcinoma in histopathological stage T3, similarly demonstrating transmural tumour growth, this time correctly suggesting stage T3.
between EUS and helical CT in the clinical staging of pancreatic cancer have yielded conflicting results. Of the potentially responsible methodological factors, blinding towards other imaging tests is a criterion rarely applied. In fact, it is quite difficult in a clinical setting to keep an examiner strictly blinded to the results of other tests. However, the same is also true of other imaging procedures—and therefore this also applies to published studies of CT and MRI in the locoregional staging of pancreatic cancer, for example. In the late 1980s and early 1990s, conventional (that is, non-helical) CT obtained accuracy rates of approximately 90% in a fair number of studies; similarly, high accuracy rates were again reached with helical CT, which was then treated as the gold standard. Routine clinical abdominal MRI became more firmly established at the end of the 1990s, MRI was found to be approximately 90% accurate also, and helical CT then suddenly dropped to low accuracy rates of approximately 75% in four comparative studies. Similar observations have been made in many other areas—for example, recently after the establishment of positron emission tomography for the differential diagnosis of pancreatic tumours. A considerable bias is therefore introduced with each new method when the examiners concerned wish to demonstrate that it is superior to previously established procedures.

These phenomena can probably be explained, at least in part, by methodological differences between the various studies. Small patient numbers, patient selection, the parameters chosen for evaluation, prospective or retrospective analysis, and blindedness in particular, are important factors, to name a few. The examiner's expectations certainly play a crucial role, and these are supported by the clinical impression and the results of other imaging tests conducted prior to the test under consideration. Our results shed light on various aspects of this question. Firstly, the accuracy of EUS (approximately 75%) in routine clinical conditions was less spectacular than previously reported, particularly in the case of oesophagogastric tumours. Secondly, when the investigator was blinded, there was a substantial reduction in the accuracy rate to approximately 50–60%, with little difference between the three types of tumours. However, adding the information from one other imaging test (endoscopy or CT) increased the accuracy again by 15–20%, to almost the same level as in routine clinical conditions. These observed effects were not evident in the case of gastric cancer, where the accuracy rates did not improve. However, the reason for this may be a systematic error in the staging of gastric cancers, as previously discussed, related to the difficult differentiation between subserosal (T2) and serosal (T3) infiltration and a tendency to overstage ulcerated and proximal gastric tumours.

Other factors that may explain the discrepancies between the fairly moderate clinical results in this study and those of previous publications are related to several types of biases. At first, publication bias, the selective reporting of studies featuring positive or extreme results, may add to this phenomenon. This has recently been debated. Secondly, there may be some case selection bias. It is well known from literature surveys that the best EUS results are usually obtained in cases of advanced cancer. In study populations in which the majority of patients have T3 cancers, the overall results are therefore likely to be better than in a study such as the present one in which T3 cancers made up less than half of the cases. For example, in a survey published in 1995, 73% of the tumours included in studies of oesophageal cancer staging using EUS were T3/T4 lesions. This effect of improving the overall results due to analysing considerably more patients with advanced cancers is also evident in our own data: the sensitivity of EUS for staging T1/T2 tumours was considerably lower in comparison with the sensitivity for correctly diagnosing more advanced tumours. A further bias may be institutional where initial reports may come from centres of excellence which do not reflect the value of the method in clinical practice—taking into account the learning curve for EUS—but there are few objective data on this in the literature.

To the best of our knowledge, this is the first study in which the accuracy of an imaging procedure such as EUS in the T staging of upper gastrointestinal tract tumours has been assessed not only in a clinical setting but also by blinding the investigator to the clinical findings and other tests, thereby ensuring a more scientific approach in order to test the “real” efficacy of EUS. Reviewing videotape sequences may be inferior to assessment at the live examination; however, well documented tape sequences were selected in which the tumour was fully visualised and examined according to a standardised approach. In addition, a potentially confounding effect of interobserver variability in explaining the divergent results obtained by different methods of data assessment is quite unlikely, in our opinion, as (at least in oesophagogastric tumours where this was tested) the same phenomenon was observed with a second investigator.

The data presented here suggest that the best methodology for evaluating imaging tests is a matter that requires debate. Strictly blinded conditions may yield the scientifically most valid data but at the same time create an artificial situation as in clinical practice the examiner always sees the patient and ideally also takes his or her history. Complementary tests may also be reviewed prior to EUS, and mistakes or complications often have to be avoided. On the other hand, in clinical conditions, these additional data obviously have a significant influence on the performance of the test under consideration (EUS in the present case) so that “true” accuracy rates cannot really be expected. We have therefore also suggested previously that a distinction needs to be made between “scientific” and “clinical” accuracy. The consensus regarding the “scientific” value of a method also has an impact on clinical case management as it is used to prevent underdiagnosis, overdiagnosis, or misdiagnosis resulting from overestimation of the effectiveness of a single method. The methodology used in studies should be made fully explicit in all cases so that the circumstances in which the respective results were achieved can be recognised.

ACKNOWLEDGEMENT
We thank M Robertson for his assistance with the English language of the manuscript.

Authors’ affiliations
A Meining, H J Dittler, A Wolf, R Lorenz, V Schusdziarra, J-R Siewert, M Classen, H Höfler, T Rösch, Departments of Internal Medicine II, Surgery, and Pathology, Klinikum rechts der Isar, Technical University of Munich, Germany

REFERENCES

To the best of our knowledge, this is the first study in which the accuracy of an imaging procedure such as EUS in the T staging of upper gastrointestinal tract tumours has been assessed not only in a clinical setting but also by blinding the investigator to the clinical findings and other tests, thereby ensuring a more scientific approach in order to test the “real” efficacy of EUS. Reviewing videotape sequences may be inferior to assessment at the live examination; however, well documented tape sequences were selected in which the tumour was fully visualised and examined according to a standardised approach. In addition, a potentially confounding effect of interobserver variability in explaining the divergent results obtained by different methods of data assessment is quite unlikely, in our opinion, as (at least in oesophagogastric tumours where this was tested) the same phenomenon was observed with a second investigator.

The data presented here suggest that the best methodology for evaluating imaging tests is a matter that requires debate. Strictly blinded conditions may yield the scientifically most valid data but at the same time create an artificial situation as in clinical practice the examiner always sees the patient and ideally also takes his or her history. Complementary tests may also be reviewed prior to EUS, and mistakes or complications often have to be avoided. On the other hand, in clinical conditions, these additional data obviously have a significant influence on the performance of the test under consideration (EUS in the present case) so that “true” accuracy rates cannot really be expected. We have therefore also suggested previously that a distinction needs to be made between “scientific” and “clinical” accuracy. The consensus regarding the “scientific” value of a method also has an impact on clinical case management as it is used to prevent underdiagnosis, overdiagnosis, or misdiagnosis resulting from overestimation of the effectiveness of a single method. The methodology used in studies should be made fully explicit in all cases so that the circumstances in which the respective results were achieved can be recognised.

ACKNOWLEDGEMENT
We thank M Robertson for his assistance with the English language of the manuscript.

Authors’ affiliations
A Meining, H J Dittler, A Wolf, R Lorenz, V Schusdziarra, J-R Siewert, M Classen, H Höfler, T Rösch, Departments of Internal Medicine II, Surgery, and Pathology, Klinikum rechts der Isar, Technical University of Munich, Germany

REFERENCES
Imaging methodology in endosonographic cancer staging


Gut through the ages

Browse the Archive

Gut online has an archive of content dating back to 1966.
Full text from 1997; abstracts from 1975; table of contents from 1966

www.gutjnl.com