Accuracy of Diagnostic Tests for *Helicobacter pylori*: A Reappraisal

Xavier Calvet,1,6,8 Jordi Sánchez-Delgado,1,6,8 Antònia Montserrat,1,6,8 Sergio Lario,1,6,8 María José Ramírez-Lázaro,1,6,8 Mariela Quesada,1,6,8 Alex Casalots,3,7 David Suárez,3,7 Rafel Campo,1,6,8 Enric Brullet,1,6,8 Félix Junquera,1,6,8 Isabel Sanfeliú,1,6,8 and Ferran Segura2,6,9,

1Digestive Diseases Department and 2Servei de Malalties Infeccioses, Hospital de Sabadell, 3Servei de Patologia and 4Laboratori de Microbiologia, UDIAT-Centre Diagnòstic, and 5Unitat d’Epidemiologia i Avaluació, Fundació Parc Taulí, Corporació Parc Taulí, 6Institut Universitari Parc Taulí, Departament de Medicina, and 7Institut Universitari Parc Taulí, Universitat Autònoma de Barcelona, 8Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Instituto de Salud Carlos III, and 9Spanish Network for the Research in Infectious Diseases (REIPI RD06/0018), Sabadell, Barcelona, Spain

**Background.** Despite many changes, no large studies comparing the different diagnostic tests for *Helicobacter pylori* have been performed in the past 10 years. In this time, monoclonal stool antigen immunoassays and in-office 13C-urea breath tests (UBTs) have appeared. The aim of this study was to evaluate the accuracy of invasive and noninvasive tests in a large series of dyspeptic patients.

**Methods.** A total of 199 dyspeptic patients who had not previously been treated for *H. pylori* infection were prospectively enrolled. Noninvasive analyses included a commercial infrared-based UBT and a commercially available stool test. Biopsy-based tests included histological examination and a rapid urease test. A patient was considered to be infected when at least 2 test results were positive. Sensitivity, specificity, positive and negative predictive values, and 95% confidence intervals were calculated. The test results were compared using the McNemar test.

**Results.** Rates of positive test results were similar (54%) for the rapid urease test, histopathological examination, and the stool test. By contrast, 75% of UBT results were positive, and the UBT was associated with a very low specificity (60%). For this reason, the delta cutoff value for the UBT was recalculated as 8.5%. Sensitivities and specificities with this new cutoff value were 95% and 100%, respectively, for the rapid urease test; 94% and 99%, respectively, for histopathological examination; 90% and 93%, respectively, for the stool test; and 90% and 90%, respectively, for the UBT.

**Conclusions.** Histological examination and rapid urease testing showed excellent diagnostic reliability. The stool test seems to be a good, noninvasive alternative to endoscopy-based tests. By contrast, the infrared-based UBT evaluated in our study showed a lower than expected performance, which was partially corrected when the cutoff value for the test was recalculated.

The methods used to diagnose *Helicobacter pylori* infection were extensively evaluated 10–15 years ago [1–5]. Since then, new developments, such as fecal tests [6–8], have appeared. Although first-generation polyclonal tests have been shown to be largely unreliable, new monoclonal tests achieve far better results [9]. Subtle changes have also been introduced in some of the classic tests. For example, one of the most reliable tests for diagnosing *H. pylori* infection, the 13C-urea breath test (UBT), is now marketed for use with a new non-dispersive, isotope-selective infrared spectroscope (NDIRS). These devices have been shown to be as reliable as isotope ratio mass spectrometers (IRMS) but are far smaller and cheaper, and they allow for in-office, near-immediate reading of results [10–12]. Some of the kits marketed in Europe for this tool, however, use a shorter time for reading than has previously been recommended (20 min, instead of 30 min) [13, 14]. In addition, unlike the kits marketed by Otsuka in the United States [15], the European kits do not include coadministration of citric acid, which is a measure that...
has been shown to increase the UBT’s sensitivity and to reduce “gray area” results [16, 17]. Validation studies to establish the cutoff value for this test were performed in Japan [18–21] but, to our knowledge, have not been reproduced in other settings.

No face-to-face comparison of these new or modified diagnostic methods has been performed to date. A reappraisal of the diagnostic tests for H. pylori infection seems to be warranted. The objective of the present study was to evaluate the effectiveness of current devices for diagnosing H. pylori infection in dyspeptic patients in a large-scale trial.

PATIENTS AND METHODS

Outpatients sent to the Endoscopy Unit of the Hospital de Sabadell (Barcelona, Spain) for dyspeptic symptoms from February 2006 through January 2008 were prospectively recruited for the study. Patients were contacted before the endoscopic examination and were asked to participate in the study. Those who agreed were instructed to avoid antisecretory drugs in the 2 weeks before the endoscopic examination. Patients who were unable to stop antisecretory drugs, those who had received antibiotics in the 4 weeks before the endoscopic examination, and those with previous treatment for H. pylori infection were excluded. Patients were asked to bring a fecal sample on the day that the endoscopic examination was to be performed. Before the endoscopic examination, the patient signed an informed consent form, and a UBT was administered. During endoscopic examination, 2 antral biopsy samples for histological examination and 1 antral biopsy sample for urease testing were obtained. Two hundred nine patients were included in the study. Ten of these patients were excluded, because either UBT test results, fecal samples, or histological examination findings were unavailable, for a variety of technical reasons. The remaining 199 patients were available for analysis. Patients’ clinical and demographic data are shown in table 1.

Diagnostic Tests

Sample collection and analysis. Fresh fecal samples were collected, aliquoted, and stored at −80°C until analysis. The tests (histological examination, rapid urease test [RUT], UBT, and the stool test) were performed by different operators who were unaware of the results of the other assessments.

Fecal test. Amplified IDEIA Hp StAR (Thermo Fisher Scientific), which is a monoclonal enzyme immunoassay, was used for the study. This stool test has shown the best diagnostic performance in previous evaluations [9, 22–27]. The test was conducted in accordance with the manufacturer’s specifications, with dual wave length readings. Samples with absorbances >0.150 were considered to be positive.

UBT. UBT was performed using UBiT 100 mg (Otsuka Pharmaceutical Europe). Determinations were performed in accordance with the manufacturer’s specifications. A basal breath sample was collected by blowing into a specifically designed bag. After this, patients drank a solution of 100 mg of 13C-labelled urea in 100 mL of water, washed their mouth out with water, and 20 min later, filled a second breath bag. Samples were immediately processed by NDIRS (POCone Infrared Spectrophotometer; Otsuka Pharmaceutical). In accordance with the manufacturer’s specifications, an increase in the 13C:12C ratio (delta 13CO2) of ≥2.5% after urea intake was considered to be indicative of H. pylori infection. Given that the 2.5% cutoff value generated an unexpected rate of false-positive results, 2 post hoc analyses were performed. First, to evaluate the performance of infrared spectroscopic examination, 120 additional patients who underwent a routine UBT with 13C urea plus citric acid in accordance with the current standard European protocol (Tau-Kit; ISomed) [16] had duplicate samples collected that were evaluated by both IRMS and NDIRS. Because the correlation of readings was near perfect (see Results), a post hoc statistical analysis was performed with the data for the 199 patients from the original study to establish the optimal cutoff value for the UBT. Final sensitivities and specificities of the different diagnostic techniques were calculated according to the modified cutoff value for this UBT.

RUT and histological analysis. RUT was performed after mucosal samples were obtained with the JATROX HP test (CHR Heim Arzneimittel) and results were read according to the manufacturer’s specifications. Samples with negative urease test results were routinely incubated overnight at 37°C. However, none of the samples with initial negative RUT results had subsequent positive results. Biopsy samples for histological examination were collected in formalin, stained with Giemsa, and evaluated by 2 pathologists who specialized in digestive diseases. The pathologists were blinded to the results of the other diagnostic tests.
Diagnostic Tests for *H. pylori* • CID 2009:48 (15 May) • 1387

Figure 1. Individual values of the 32 false-positive urea breath test (UBT) results, with positivity defined according to the 2.5% delta cutoff value recommended by the manufacturer. Except for 3 outlier results, most of the false-positive test results had low delta values (2.5%–12%).

Statistical Methods

**Sample size calculation.** Assuming a prevalence of *H. pylori* infection of 75% in the population with dyspepsia that was evaluated [28, 29], a sample size of 203 patients was necessary to obtain an estimation of sensitivity with a 95% confidence interval of 0.05 and a confidence level of 0.9.

**Statistical analysis.** *H. pylori* infection status was determined on the basis of the combined results of the RUT, histological examination, UBT, and fecal tests. Patients with ≥2 positive test results were considered to be infected. Sensitivity, specificity, and positive and negative predictive values were calculated for each diagnostic technique and expressed as percentages and 95% confidence intervals. Given the unexpectedly high rate of false-positive UBT results obtained when using the cutoff value provided by the manufacturer, a post hoc evaluation was performed: a receiver operating characteristic curve was drawn to determine the cutoff value that demonstrated the best sensitivity and specificity. Linear correlation was used to compare values of the readings from NDIRS and IRMS, and mean values were compared by Student’s *t* test. The McNemar test was used to compare the sensitivity and specificity of the different tests [30]. To correct for multiple comparisons, only *P* values of ≤.01 were considered to be significant. Quantitative variables are given as mean values ± standard deviations. All calculations were performed using SPSS, version 15.0 for Windows (SPSS). The study was performed in compliance with the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) recommendations [31].

RESULTS

Among 199 patients, 107 (54%) had positive RUT results, 107 (54%) had positive histopathological findings, 108 (54%) had positive stool test results, and 149 (75%) had positive UBT results. In accordance with the preestablished gold standard, 118 patients (59%) with ≥2 positive test results were considered to be infected. With use of the recommended delta value of 2.5%, the UBT had a sensitivity of 99.2% but a specificity of only 60.5%, with 32 false-positive results. Most of these false-positive results were in a “gray area” of delta values, between 2.5% and 12%, with a median value of 5.4% and an interquartile range of ±4.7% (figure 1). In this initial analysis, the sensitivities and specificities of the remaining tests were 91% and 100%, respectively, for histological analysis; 91% and 100%, respectively, for RUT; and 90% and 98%, respectively, for the stool test. No statistically significant differences were observed between tests, except for the UBT, which demonstrated statistically significantly less specificity than the other 3 tests (*P* < .01, by McNemar test). Figure 2 shows a STARD flow diagram for calculations performed with the manufacturer’s recommended cutoff for UBT.

**Recalculating a cutoff value for UBT.** To evaluate whether the high rate of false-positive results was attributable to the infrared reading device, dual readings by means of IRMS and NDIRS were performed for 120 additional patients. Comparison of the IRMS and NDIRS showed a near perfect correlation (r = 0.992). The infrared reader led, however, to a minor but statistically significant decrease in the mean estimation of delta values (13.4 ± 24 vs. 14.4 ± 24; *P* < .001, by Student’s *t* test). A receiver operating characteristic curve for UBT was drawn (figure 3). In view of the receiver operating characteristic curve data, a cutoff value of 8.5% was chosen to maintain high sensitivity and acceptable specificity.

**Sensitivity, specificity, and positive and negative predictive values of the different techniques.** With the new cutoff value for UBT, 113 patients (56.8%) had ≥2 positive test results. A STARD flow diagram of these new calculations is shown in

Figure 2. Standards for the Reporting of Diagnostic Accuracy Studies diagram depicting the flow of patients and the results of the analysis. Values are calculated for a urea breath test (UBT) cutoff value of 2.5%. Note the high rate of false-positive results of UBT, compared with the rates for the other tests. GS, gold standard; RUT, rapid urease test.
Figure 3. Receiver operating characteristic curve for the UBiT 100 mg test (Otsuka Pharmaceutical Europe). The area under the curve is 0.948.

Figure 4. Standards for the Reporting of Diagnostic Accuracy Studies diagram depicting the flow of patients and the results of the analysis. Values are calculated for a urea breath test (UBT) delta cutoff value of 8.5%. GS, gold standard; RUT, rapid urease test.

DISCUSSION

The most striking finding of the study was the unexpectedly low reliability of UBT. If adequately performed, UBT has been shown to be as reliable as histological examination. The poor performance of the test in our series may be attributable to 2 main reasons. First, the low cutoff value was established in a series of only a few individuals and, to our knowledge, has not been retested: there are no published validations of this particular kit in the medical literature outside of Japan and, indeed, very few there [18–21]. In the package insert of the test approved by the US Food and Drug Administration, data are reported for 499 duodenal ulcer patients [15]. Although the specificity of the test was 89% with use of a delta cutoff value of 2.4%, specificity was calculated on the basis of data from only 30 patients with H. pylori–negative duodenal ulcers. Therefore, the data on the adequacy of the cutoff value established for Otsuka’s Breathtek and Meretek UBT kits commercialized in the United States are also limited. In addition, both the test (which includes the preadministration of citric acid) and the population (patients with duodenal ulcers instead of patients with uninvestigated dyspepsia) differed from those in the present study.

In our study, establishing a new, higher cutoff value for the test considerably improved the reliability of the technique, although sensitivity and specificity both remained ~90%, which are relatively low values for UBT. This may be related to the second important point: unlike the kits marketed by Otsuka in the United States [15], European kits by this manufacturer do not use concomitant citric acid [13, 14]. Most studies evaluating the need for citric acid in UBT showed higher delta values with citric acid when compared with other test meals or no test meals [16, 17, 32, 33], although this increase was not statistically significant in some Chinese studies [34]. Citric acid is expected to increase delta values in infected patients and not change delta values in uninfected ones. Adding citric acid may, therefore, well increase the discriminative capacity of the test, reducing the number of patients with “gray area” delta values [17, 33]. However, the improvement is expected to be modest. As our study shows, UBT is a good test even without citric acid, reaching sensitivities and specificities of ~90%. Citric
Therefore, that in addition to being evaluated as positive or which is near the cutoff for positivity. It should be stressed, results in this series had a range of delta values of 8.5%–12%, not very high or for use as an in-office test [36].

Laboratories in which the daily number of assays performed is operating characteristics make NDIRS particularly suitable for ing time is longer, and NDIRS can process only a few breath and economic: NDIRS devices are much cheaper, but the read-

Differences between IRMS and NDIRS are basically logistical excellent correlations between IRMS and NDIRS [10, 11, 35]. In fact, the studies comparing citric acid versus no test meal or citric acid versus other test meals usually show a slight improvement in UBT accuracy with the use of citric acid. This improvement is rarely statistically significant, because the sample size needed to demonstrate a statistically significant improvement of sensitivity or specificity from 90% to 95% is extremely large, and most of the studies on the topic have been rather small [16, 17, 32–34]. In addition, adding citric acid to UBT is economical, simple, and safe, thus justifying its generalized use [16, 17, 32–34]. In fact, the European Standard Protocol includes citric acid with the kit [16]. Taking all of the data into account, our results strongly suggest that the UBT should be modified by including citric acid coadministration to improve its diagnostic accuracy.

What did not seem to influence the results of UBT is the method used to read the test. IRMS and NDIRS obtained very similar values when dual readings were performed. This finding is in accordance with other reports in the literature: several noncomparative and comparative studies have demonstrated excellent correlations between IRMS and NDIRS [10, 11, 35]. Differences between IRMS and NDIRS are basically logistical and economic: NDIRS devices are much cheaper, but the reading time is longer, and NDIRS can process only a few breath samples at the same time. NDIRS also requires the connection of large breath bags to the spectrometer for measurement, which makes it more difficult to store and transport large volumes of breath samples to a measuring laboratory. All of these operating characteristics make NDIRS particularly suitable for laboratories in which the daily number of assays performed is not very high or for use as an in-office test [36].

A final point regarding UBT is that most false-positive test results in this series had a range of delta values of 8.5%–12%, which is near the cutoff for positivity. It should be stressed, therefore, that in addition to being evaluated as positive or negative, UBT results should also be read quantitatively. The study shows that the positive predictive value of delta values >12% is practically 100%. Therefore, a borderline UBT and a strongly positive test result are not the same in terms of confirming the presence of H. pylori infection, a finding which we already observed in previous studies [37] and which should be taken into account in clinical practice.

The present study also shows that biopsy tests continue to be an uncontroversial standard for diagnosing H. pylori infection. Although slightly dependent on the observer and the staining method, biopsy tests achieved sensitivities and specificities >95% [38]. In addition, the combination of RUT and histological examination provided a very high diagnostic yield. The combination of the 2 tests seems sufficient either to confirm or to rule out H. pylori infection in clinical practice, at least for previously untreated patients with uninvestigated dyspepsia, such as those included in the present series.

Regarding the monoclonal enzyme-linked immunosorbent assay Amplified IDEIA Hp StAR test, the results showed that, although it was somewhat less sensitive than histological examination, the test achieved sensitivities and specificities >90%, making it suitable for use in clinical practice. However, this result applies only to this particular kit. The efficacy of stool tests for detecting H. pylori infection depends greatly on the antigen selected for detection. Indeed, it has been shown that polyclonal antibody tests, which have different antigenic compositions depending on the batch used in the test, show very large intratest variability [39–41] and poor reliability [9, 42]. Overall, the results are far less reliable than those of monoclonal antibody stool tests [9]. In addition, because not all monoclonal antibody tests detect the same antigen, genetic variations of H. pylori strains could lead to geographical variations in diagnostic efficacy. Their usefulness should, therefore, be tested locally. Amplified IDEIA Hp StAR, however, has generally shown good results in the various settings in which it has been evaluated [24–27, 43–50]. Finally, the method of detection of the antigen is also important: immunoassays are more reliable than

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>Positive predictive value, % (95% CI)</th>
<th>Negative predictive value, % (95% CI)</th>
<th>Positive likelihood ratio</th>
<th>Negative likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid urease test</td>
<td>94.7 (88–98)</td>
<td>100 (95–100)</td>
<td>100 (95–100)</td>
<td>93.5 (86–97)</td>
<td>0.108</td>
<td>0.053</td>
</tr>
<tr>
<td>Histopathological examination</td>
<td>93.8 (87–97)</td>
<td>98.8 (93–100)</td>
<td>99.1 (94–100)</td>
<td>92.4 (85–97)</td>
<td>78.2</td>
<td>0.063</td>
</tr>
<tr>
<td>Urea breath test&lt;sup&gt;a&lt;/sup&gt;</td>
<td>90.3 (83–95)</td>
<td>89.5 (81–95)</td>
<td>91.9 (83–95)</td>
<td>87.5 (78–93)</td>
<td>8.6</td>
<td>0.108</td>
</tr>
<tr>
<td>Amplified IDEIA Hp StAR (Thermo Fisher Scientific)</td>
<td>90.3 (83–95)</td>
<td>93 (84–97)</td>
<td>94.4 (88–98)</td>
<td>87.9 (79–94)</td>
<td>12.9</td>
<td>0.104</td>
</tr>
</tbody>
</table>

<sup>a</sup> These values are calculated for a delta cutoff value of 8.5%.
in-office, immunochromatographic tests. Moreover, the quality and reliability of immunochromatographic methods vary widely.

Strengths of this study are that it includes a large sample of ambulatory, unselected dyspeptic patients and that all of the tests were prospectively performed within a short time interval. The sample included is large enough to provide data on the reliability of the different tests with close confidence intervals. The possibility of an error when performing the UBT was also evaluated: the process of collecting the samples and performing the analysis was reviewed jointly with the manufacturers’ technicians to rule out an error in obtaining or processing the samples. In addition, we performed a second study to compare the mass spectrometer with the infrared UBT reader and observed a strong correlation between tests, thus suggesting that both the sample collection and the reading of the samples were adequately performed. A limitation of the study is that we could not fully analyze the reasons for the poor performance of the UBT. The likely efficacy of adding citric acid to improve test results should be tested in additional studies.

The use of a post hoc cutoff value for the UBT instead of the manufacturer’s established cutoff value could be viewed as a limitation. To avoid this possible drawback, the study clearly separates the data on the accuracy of UBT according to the delta value employed. In addition, the post hoc cutoff calculation was necessary, because it is informative and useful for the reader and gives an indication of the most appropriate cutoff value in clinical practice. The analysis showed that, even with the new cutoff value, the test did not achieve the sensitivity and specificity of 95% that could be expected for the UBT. These findings highlight the need to use citric acid to achieve optimal results with the UBT.

In conclusion, our study shows that histological examination and RUT remain the uncontested gold standard for diagnosing *H. pylori* infection. Amplified IDEIA Hp StAR can be considered as a noninvasive first-line routine diagnostic test. In contrast, at least some of the IRMS-based UBTs display a low diagnostic efficacy, which may be attributable to recent modifications in the technique. Because the UBIT 100 mg is currently the most frequently used UBT in Spain, and because the kit has been approved by the European Medicaments Agency and is soon to be launched in many other European countries, the findings of the present study must be urgently confirmed or disproved. Use of the test with the current cutoff value could produce a high rate of false-positive results and lead to unnecessary treatment and associated costs. Additional studies are necessary to set a new cutoff value for the test and to evaluate the need for citric acid pretreatment to improve the diagnostic reliability of UBT.

**Acknowledgments**

We thank Neus Mateo and Laura Moreno, for their help in collecting samples and performing the breath test, and Michael Maudsley, for his help with the English language version of this article.

**Financial support.** Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas is funded by the Instituto de Salud Carlos III. The study was also supported by grants from the Instituto de Salud Carlos III (PI 05/1157 and PI 05/0664) and from the Societat Catalana de Digestologia. X.C. has a personal research grant from the Programa de Intensificación en Investigación del Instituto de Salud Carlos III.

**Potential conflicts of interest.** All authors: no conflicts.

**References**