

# Short-Term Treatment with Proton-Pump Inhibitors as a Test for Gastroesophageal Reflux Disease

## A Meta-Analysis of Diagnostic Test Characteristics

Mattijs E. Numans, MD, PhD; Joseph Lau, MD; Niek J. de Wit, MD, PhD; and Peter A. Bonis, MD

**Background:** A response to proton-pump inhibitors (PPIs) is commonly considered to support the diagnosis of gastroesophageal reflux disease (GERD). However, the accuracy of this diagnostic strategy has not been well established.

**Objective:** To estimate the diagnostic test characteristics of successful PPI treatment with objective measures of GERD by performing a meta-analysis based on the published literature.

**Data Sources:** English-language studies were identified by searching the Cochrane Clinical Trial Register and MEDLINE from 1 January 1980 through 1 July 2003.

**Study Selection:** Studies in which the clinical response to a short course (1 to 4 weeks) of normal- or high-dose PPI therapy could be compared with objective measures of GERD (24-hour pH monitoring, endoscopy findings, symptom questionnaires).

**Data Extraction:** Studies were screened for inclusion by 1 author. Final decisions on exclusion were made by consensus with 2 of the other authors. Two investigators independently extracted the data. Information extracted included patient characteristics, study design, setting, specific type and dose of medication, duration of treatment, and definitions of outcomes.

**Data Synthesis:** Sensitivity and specificity were determined by comparing a clinical response to PPIs with objective measures for GERD. The summary receiver-operating characteristic curve method was used to summarize test characteristics across studies. Sensitivity and specificity were also combined independently by using a random-effects model. Fifteen studies met the inclusion criteria and provided sufficient data. With 24-hour pH monitoring as the reference standard, the positive likelihood ratio ranged from 1.63 to 1.87, and combined estimates of sensitivity and specificity were 0.78 (95% CI, 0.66 to 0.86) and 0.54 (CI, 0.44 to 0.65), respectively. These values were lower with the other reference standards.

**Limitations:** Data were insufficient to determine the effect of various doses of PPIs and duration of therapy on test characteristics.

**Conclusion:** Successful short-term treatment with a PPI in patients suspected of having GERD does not confidently establish the diagnosis when GERD is defined by currently accepted reference standards.

*Ann Intern Med.* 2004;140:518-527.

[www.annals.org](http://www.annals.org)

For author affiliations, see end of text.

Patients with symptoms suggestive of uncomplicated gastroesophageal reflux disease (GERD) are frequently treated empirically with lifestyle modifications and acid suppressive medications (1, 2), a strategy endorsed by several major medical societies (3, 4). Proton-pump inhibitors (PPIs) are often selected as first-line therapy because they are more effective than other available treatments for GERD (5). Furthermore, a rapid symptomatic response to normal-dose PPIs in patients with a presumptive diagnosis of GERD is commonly considered to validate the diagnosis (the "PPI test") (6). However, the accuracy of a symptomatic response to PPIs compared with objective measures of GERD (such as ambulatory pH monitoring) is unclear. Studies addressing this issue have produced variable estimates of test accuracy (7–11). The differences may be due in part to several factors, such as the dose of the PPI used, population-specific features, specific end points, and the reference standard measures used to define GERD.

A more thorough understanding of the characteristics of the PPI test is needed, especially from a primary care perspective, because the diagnosis of GERD can have significant clinical and economic implications. False-positive results on the PPI test, for example, may result in long-term administration of PPIs in patients who do not have GERD. In contrast, false-negative results may lead to under-treatment or incorrect diagnosis in patients with GERD.

Our study sought to better define sensitivity, specificity, and predictive values of the PPI test for GERD while delineating factors that have a bearing on test accuracy. We performed a systematic review and meta-analysis of diagnostic test characteristics by using responses to PPI treatment as a test to compare with several objective measures of GERD.

## METHODS

### Search Strategy

Many symptoms have been attributed to gastroesophageal reflux. As a general rule, the term *GERD* is applied to patients with symptoms suggestive of reflux or complications of reflux but not necessarily with esophageal inflammation (12). We sought studies that enrolled patients on the basis of symptoms that were suggestive of GERD. Studies were identified by searching the MEDLINE database (1 January 1980 through 1 July 2003) and the Cochrane Controlled Trial Register. Search terms were *omeprazole*, *lansoprazole*, *pantoprazole*, *rabeprazole*, and *esomeprazole*, expanded to *randomized controlled trials*, *random allocation*, *double blind*, *single blind*, *comparative studies*, *esophageal motility disorders*, or *esophagitis*. Additional search terms were *GERD* or *GORD*, expanded to *diagnosis*, *reproducibility of results*, *false negative reactions*, *false positive*

reactions, logistic models, regression analysis, predictive value, sensitivity and specificity, accuracy, screening, and likelihood ratio (13).

The searches were limited to English-language studies performed in adults (age  $\geq 18$  years). We reviewed the bibliographies of relevant studies to search for additional eligible studies. Only data accessible in peer-reviewed journals were included to minimize potential sources of bias and inaccuracy (14).

### Inclusion Criteria

We included studies that assessed a symptomatic response in adults with a presumptive diagnosis of GERD (on the basis of presenting symptoms and history) who were treated with a PPI and who also underwent objective testing for GERD with 24-hour pH monitoring or endoscopy or both. We also evaluated studies that used a structured symptom scoring system as the reference standard, a method that has been proposed to be useful for discriminating patients with GERD from those with other causes of symptoms (15–17). These studies focused mostly on patients with dyspepsia, in whom the symptom scores were used to distinguish patients with “reflux-like” symptoms from those with other symptom characteristics. Such studies were included if the outcomes after PPI treatment in patients with GERD-like symptoms could be compared with the outcomes of patients with other symptom characteristics and if patients had been included irrespective of their symptom characteristics.

### Exclusion Criteria

Studies were excluded if they focused only on children (age  $< 18$  years), if they included only patients with complications caused by GERD (such as esophageal strictures, severe esophagitis, Barrett esophagus), patients with alarm symptoms (such as anemia or dysphagia), patients with extraesophageal symptoms of GERD (such as asthma or laryngitis), patients suspected of having cardiovascular disease, or patients in whom a symptomatic response to the PPI could not be correlated with any of the objective tests for GERD (24-hour pH monitoring, upper endoscopy, or a symptom score).

### Data Extraction

Studies were screened for inclusion by one author. Final decisions on exclusion were made by consensus with 2 of the other authors. Two investigators independently extracted the data. Information extracted included patient characteristics, study design, setting, specific type and dose of medication, duration of treatment, and definitions of outcomes. Numbers were extracted directly from the tables or derived from percentages if only the total number of patients was available. Data were recorded separately for the appropriate treatment group in studies with a randomized or crossover design. Discrepancies were resolved by discussion until consensus was achieved for all data.

### Context

Does symptomatic response to a proton-pump inhibitor (PPI) diagnose gastroesophageal reflux disease (GERD)?

### Contribution

This meta-analysis included 15 studies that compared clinical response to a short course of a PPI with an objective measure of GERD, such as 24-hour pH monitoring. The positive likelihood of a symptomatic response detecting GERD ranged from 1.63 to 1.87. Sensitivity was 0.78 (95% CI, 0.66 to 0.86) and specificity was 0.54 (CI, 0.44 to 0.65).

### Implications

Symptomatic response to short-term treatment with a PPI does not confidently diagnose GERD.

—The Editors

### Quality Assessment of Studies

Studies included in the meta-analysis were assessed for quality, as suggested by a previously reported standard (18). We particularly focused on whether the authors clearly defined normal and abnormal results according to the reference standard, whether the study sample included those with and without GERD according to the reference standard, whether the authors adequately described how the reference tests were performed, whether test results were interpreted appropriately, whether the report presented the results with sufficient detail so that the study could be replicated, and whether the outcome measures and objective measure of GERD were consistent with accepted standards for the evaluation of GERD (12). Our study selection criteria required fulfillment of most of the quality standards.

### Definition of PPI Treatment Success as a Diagnostic Test

The definition of a symptomatic response to a PPI was based on criteria defined in the individual reports. Whenever possible, we chose definitions that would reasonably be interpreted as representing success in clinical practice (Table 1). As a general rule, we considered “complete relief of heartburn” as representing success because it represents a reasonable end point in clinical practice.

### Reference Standards for GERD

The following reference standards for GERD were used in this study, all of which were based on commonly accepted measures (12).

### Ambulatory Monitoring

Ambulatory esophageal pH monitoring is generally considered to provide the most objective measurement of pathologic reflux. Results were considered to be abnormal if an esophageal pH less than 4.0 was recorded during more than 4% of the time during a 24-hour period. This definition of esophageal pH is consistent with widely accepted standards (32).

Some investigators have used a “symptom index” in which symptoms are correlated to episodes of esophageal acidification (pH < 4.0) while patients are in the erect or supine position. A symptom index greater than 50% is considered to be abnormal (10, 23). A related measure (“symptom associated probability”) has also been described (33, 34).

Some patients may have increased sensitivity to esophageal acid exposure, thereby developing typical symptoms of GERD without abnormal esophageal acid exposure, as assessed by ambulatory pH monitoring. The clinical response to acid suppression in such patients is similar to that in patients without esophageal hypersensitivity (5). As a result, reference values and interpretation of ambulatory esophageal pH monitoring continue to undergo evaluation.

### Upper Endoscopy

Gastroesophageal reflux disease is also suggested by findings on upper endoscopy (esophagogastrodueno-

scopy [EGD]). Esophageal erosions, ulcers, Barrett esophagus, or strictures suggest the presence of GERD. However, results are frequently normal in patients with symptoms of GERD and abnormal esophageal acid exposure (32).

Gastroesophageal reflux disease was considered to be present in studies reporting esophagitis according to one of the commonly used classification systems, such as the Savary, Los Angeles, Berstad, or Hetzel–Dent grading systems (30, 31, 35, 36). Endoscopy was considered to be diagnostic of GERD in patients with supporting clinical features who had at least grade I esophagitis in one of the scoring systems. Because our study focused on the diagnosis of GERD that might safely be treated empirically, studies that focused exclusively on patients with severe or complicated esophagitis (such as Savary III or IV or Los Angeles C or D) after selection with endoscopy were excluded. Once recognized, this group of patients not only needs further investigation but might also be most likely to

**Table 1. Studies of Patients with Symptoms Suggestive of Gastroesophageal Reflux Disease in Whom the Study Outcome Is Symptom Relief with Short-Term Normal- or High-Dose Proton-Pump Inhibitor Treatment\***

Study, Country (Reference)	Study Setting	Inclusion Criteria†	Design	PPI Test
Bate et al., United Kingdom (11)	Secondary/specialist care		Open-label trial	Omeprazole, 40 mg
Brun and Sorngard, Sweden (19)	Primary care		Open-label trial	Omeprazole, 40 mg
Carlsson et al., Sweden (20)	Primary care	Carlsson–Dent score >3 (15)	RCT	Omeprazole, 20 mg
Dupas et al., France (21)	Secondary/specialist care	Esophagitis Savary II	RCT	Pantoprazole, 40 mg Lansoprazole, 30 mg
Farup et al., Norway (22)	Primary care	Functional dyspepsia	RCT	Omeprazole, 20 mg
Fass et al., United States (10)	Secondary/specialist care		RCCT§	Omeprazole, 60 mg
Fass et al., United States (23)	Secondary/specialist care	Esophagitis Hetzel–Dent score >1	RCCT§	Omeprazole, 60 mg
Galmiche et al., France (24)	Secondary/specialist care		RCT	Omeprazole, 20 mg
Hatlebakk et al., Norway (25)	Primary care		RCT	Omeprazole, 20 mg
Johnsson et al., Sweden (26)	Secondary/specialist care		RCT	Omeprazole, 40 mg
Juul-Hansen et al., Norway (9)	Secondary/specialist care	Normal endoscopy findings	RCCT§	Lansoprazole, 60 mg
Lewin van den Broek et al., the Netherlands (27)	Primary care	Uninvestigated dyspepsia	RCT	Omeprazole, 20 mg
Schenk et al., the Netherlands (8)	Secondary/specialist care		RCT	Omeprazole, 40 mg
Talley et al., Australia (28)	Primary care	Functional dyspepsia	RCT	Omeprazole, 20 mg
Venables et al., United Kingdom (29)	Primary care		RCT	Omeprazole, 20 mg

\* EGD = esophagogastroduodenoscopy; GERD = gastroesophageal reflux disease; NA = not available; PPI = proton-pump inhibitor; RCCT = randomized, controlled crossover trial; RCT = randomized, controlled trial.

† Typical symptoms of gastroesophageal disease are defined as predominantly, but not exclusively, reflux-like symptoms (heartburn or acid regurgitation).

‡ Composite symptom score (DeMeester) of more than 4 is accepted as typical GERD (17).

§ RCCTs have data from the first treatment period available.

|| Utrecht dyspepsia scoring lists (16, 27).

respond clinically to PPI treatment because the cause of symptoms is probably related to acid.

The sensitivity of endoscopy can be increased when esophageal biopsy specimens are obtained. Histologic characteristics suggesting GERD include the presence of eosinophils, thickening of the basal cell layer, and elongation of the papillae of the epithelium (37). However, histologic assessment was not performed in any of the included studies and is uncommonly performed in clinical practice in patients who have a normal-appearing esophagus.

**Symptom Scores**

The cardinal clinical features of GERD (heartburn and regurgitation) are not always present in all patients who ultimately receive a diagnosis of GERD. In clinical practice, GERD must be frequently distinguished from other causes of upper abdominal or chest discomfort. However, the ability to discriminate GERD from other disorders is not always straightforward because symptoms may overlap, leading to many false-positive and false-negative diagnoses (38).

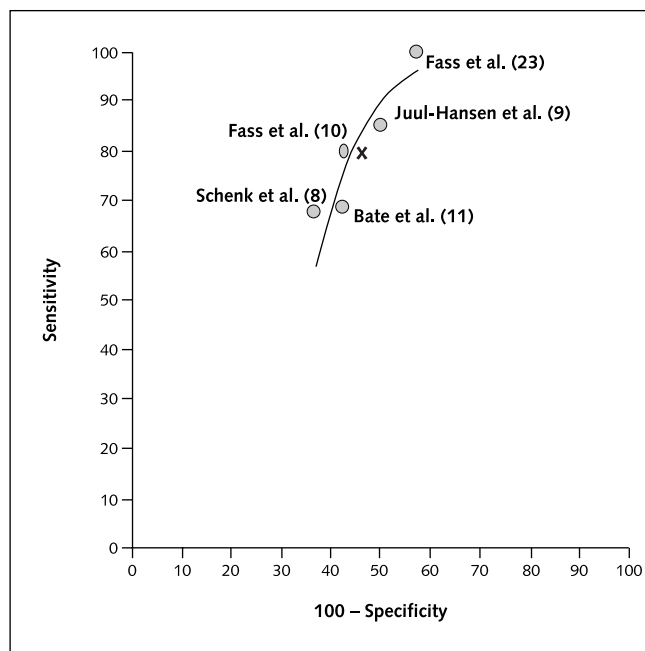
Several authors have proposed structured scoring systems based on clinical features that can be used to discriminate patients with GERD from those with other causes of symptoms. One is based on expert opinion and weighs the value of symptoms scored on a patient-administered questionnaire (15), another is based on translation of logistic regression analysis of endoscopy results into a physician-administered scoring list (16), and another is based on a mathematically reduced interpretation of symptoms asked during history taking (17). The performance characteristics of these systems have not been validated extensively, and they have not been widely adopted. In addition, the relationship between any scoring system and the PPI test has not been well studied.

To better understand the relationship between symptom score and treatment response, we considered GERD to be present when the study authors identified patients with GERD on the basis of predetermined clinical criteria (19, 28) or when symptoms had reached an adequate

Table 1—Continued

Duration	Evaluable Patients	Men	Average Age	Duration of Symptoms	Reference Standard for GERD	Population of Patients with Symptoms Suggestive of GERD at Inclusion	Prevalence of Esophagitis	Outcome Measure Extracted
<i>wk</i>	<i>n</i>	%	<i>y</i>	<i>mo</i>		%		
2	58	55.1	47.4	Average	pH <4.0 during at least 4% of monitoring time	100	51	Complete relief or >50% reduction
1	362	NA	NA	NA	Score is predominant heartburn	100	NA	Main symptoms >45% improved
4	225	50.1	50	>3	Mucosal breaks on EGD	100	61	Complete symptom relief
2	362	74	54	NA (12% first episode)	Composite symptom score >4 (15)‡	100	100	Complete symptom relief
4	14	50	50	>1	Score is heartburn or acid regurgitation	57	0	Sufficient relief: yes/no
1	42	76.2	55.4	>3	pH <4.0 during >4.2% of monitoring time	100	50	Symptom score improved ≥50%
1	35	94.3	55.4	>3	pH <4.0 during >4.2% of monitoring time	100	100	Symptom score improved ≥50%
4	141	44.7	50	>3 (26% <1 y)	Erosions in EGD; not circumferential	100	26	Symptoms decreased to 1 d/wk; less than mild
4	161	57	49	>3	Esophagitis grade 1 (Berstad [30])	100	48	Symptoms decreased to 1 d/wk; less than mild
1	80	NA	NA	>6	Savary II and III or pH <4.0 during >4% of time	100	58	Symptom score improved
5/7	56	31.3	54	>2	pH <4.0 during >4.2% of monitoring time	100	0	Antacid use decreased >75%
2	84	44.2	43.5	>2 wk (32% first episode)	Score >1 on prediction rule for GERD	49	NA	No strategy failure
2	41	39	49	NA	pH <4.0 during >4% of time and Savary I (31)	100	37	>50% reduction
4	413	39.9	42	>1 (37% <1 y)	Score is predominant heartburn or acid regurgitation	12	NA	Complete symptom relief
4	330	52	51	>3 (26% <1 y)	Erosive esophagitis (Savary II and III) (31)	100	31	Symptoms decrease to 1 d/wk; less than mild

**Figure 1.** Summary receiver-operating characteristic curve analysis of the proton-pump inhibitor test with abnormal 24-hour pH monitoring as the reference standard ( $n = 232$ ).



threshold to diagnose GERD on the basis of symptom score results (20, 21, 27) (Table 2).

### Diagnostic Test Characteristics

Sensitivity, specificity, predictive values, and positive likelihood ratios for each study were calculated from the original data by comparing the results of objective testing with proportions of successful short-term treatment with PPIs in subgroups. The likelihood ratio represents a measure of the odds of having a disease relative to the prior probability of the disease. An advantage of likelihood ratios (compared with predictive values) is that they are relatively independent of disease prevalence. The positive likelihood ratio in our analysis represents the probability that a patient with GERD (defined by the reference standard) will have a positive response to a PPI challenge divided by the probability that a patient without GERD will have a positive response to a PPI challenge. Thus, the higher the positive likelihood ratio, the better discriminant ability of the PPI test (a perfect test would have a positive likelihood ratio equal to infinity).

### Summarized Test Characteristics and Summary Receiver-Operating Characteristic Curve Analysis

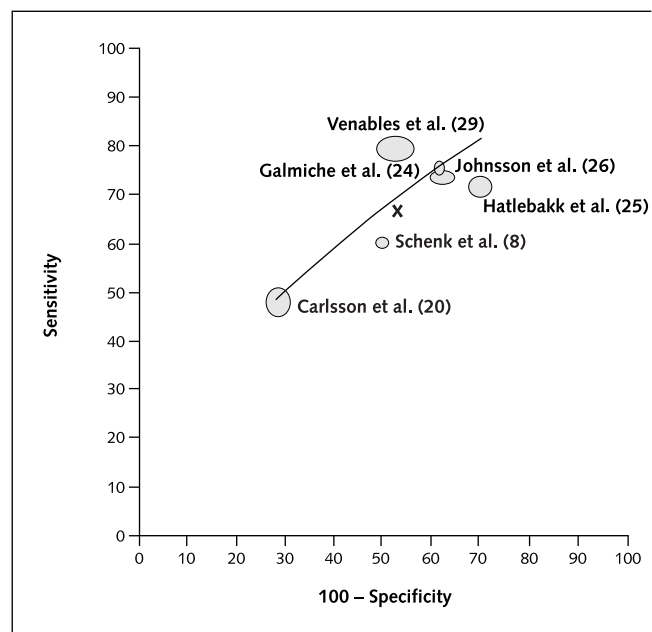
We used 2 methods to summarize the data. The summary receiver-operating characteristic (SROC) curve method was used to depict the trade-off between sensitivity and specificity among different studies with different characteristics and thresholds (Figures 1 and 2) (39, 40). An SROC curve plots the true-positive rate (sensitivity) on the y-axis against the false-positive rate ( $1 - \text{sensitivity}$ ) on the x-axis for each study. For the PPI test, there is no simple

meaning for threshold. We hypothesized that a threshold effect might find its origin in variation of setting, outcome definition, PPI dosage, sharpness of the definition of GERD, or a combination of these factors. Studies were combined by using the SROC method provided that the definitions used for the symptomatic response as well as definitions of the reference test were comparable and all numbers needed for analysis were readily available or easily derived. We made no assumption that the SROC curve is symmetrical about the axis where sensitivity is equal to specificity. As a complementary method to summarize the data, we independently combined the sensitivity and specificity of PPI treatment success for GERD across the studies by using a random-effects model (Table 3). These estimates are empirically useful as an approximation of the overall test sensitivity and specificity, although they do not provide information about the threshold effect. To ensure that the pooled CI is bounded between 0% and 100%, the logits of the proportions (that is, sensitivity or specificity) were combined by using a random-effects model. The results were then transformed back as standard proportions along with their CIs. This approach also tends to underestimate the test performance as compared with the SROC method (41, 42). All calculations were made by using MetaTest software as described previously (43, 44).

### Role of the Funding Sources

The funding sources did not control or influence the content of the research or this paper and had no role in the decision to submit the manuscript for publication.

**Figure 2.** Summary receiver-operating characteristic curve analysis of the proton-pump inhibitor test with relevant esophagitis found on esophagogastroduodenoscopy as the reference standard ( $n = 978$ ).



**Table 2. Diagnostic Test Characteristics of Successful Proton-Pump Inhibitor Treatment with the Reference Test Gastroesophageal Reflux Disease Diagnosed on the Basis of 5 Different Structured Questionnaires (n = 1235)**

Structured Questionnaire (Reference)	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Positive Likelihood Ratio
Brun and Sorngard (19)	0.73	0.29	0.81	0.20	1.03
Farup et al. (22)	0.38	0.17	0.38	0.17	0.45
Lewin et al. (27)	0.90	0.09	0.49	0.50	0.99
Talley et al. (28)	0.54	0.64	0.17	0.91	1.50
Dupas et al. (21)	0.69	0.21	0.22	0.67	0.87

## RESULTS

### Study Search Flow

Our search strategy identified 5771 studies, of which 1144 were treatment studies with one or more PPI (that is, omeprazole, esomeprazole, lansoprazole, pantoprazole, or rabeprazole) (Figure 3). Most of these focused on treatment of peptic ulcers or *Helicobacter pylori* and were therefore excluded. Of 136 potentially eligible studies, 119 were excluded because they reported insufficient data with which to calculate sensitivity and specificity. In particular, 51 studies reported outcomes in subgroups of patients with different symptoms but gave no numbers or percentages, whereas 25 gave numbers without correlation to specific symptoms and 12 reported data only on treatment success in patients with GERD without providing data on treatment success in patients without GERD. One study provided results for sensitivity only (7). Data from the most recent study on this subject could not be used because relevant information on patients who withdrew were not provided (34). Fifteen studies reported sufficient data to be included in a meta-analysis (Table 1).

### Study Characteristics and Quality

The primary goal of 6 of the studies was the investigation of the PPI test for diagnosing GERD (8–11, 23, 26). The others were randomized, controlled trials of the treatment of GERD that reported adequate data on treatment success in relevant subgroups. Seven studies enrolled patients seen in primary care settings, and 8 enrolled patients who had been referred to a gastroenterologist. Three studies included patients with dyspepsia in which the results of patients with GERD-like symptoms could be compared with those of patients with other symptom characteristics (22, 27, 28). Two studies included patients with erosive GERD as assessed on endoscopy (21, 23). One study used a GERD score as an additional inclusion criterion (20). Endoscopy was performed in 11 studies, and 24-hour pH monitoring was performed in 7. Outcomes were assessed after 5 days to 4 weeks in all studies. The doses of the different PPIs used varied. However, all doses would have been expected to achieve a substantial reduction in gastric acid secretion and were at least as high as would typically be used for an empirical PPI trial in clinical practice (Table 1).

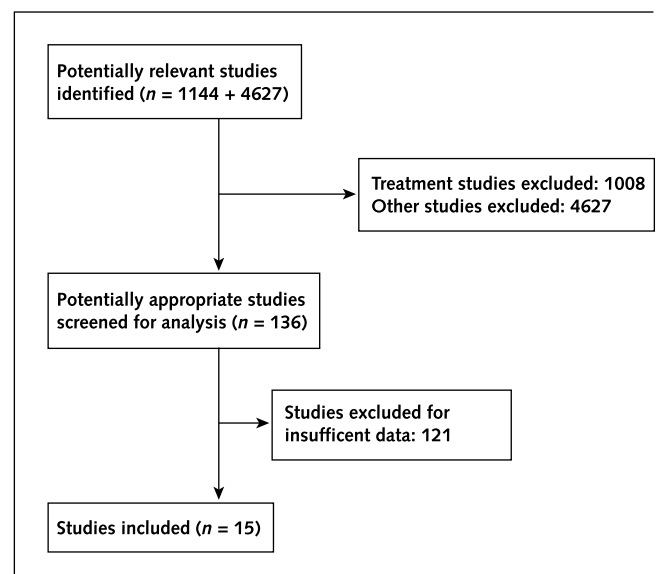
The definitions of clinical success varied across studies,

although most would have been reasonably interpreted as representing substantial improvement in clinical practice. Treatment success was defined as a reduction in symptoms to zero in 3 studies (20, 21, 28); as a reduction in symptoms to “sufficient” in 1 study (22); as a reduction in symptoms to less than 50% of baseline in 5 studies (8, 10, 11, 19, 23); as a reduction in symptoms with 1 step on the Likert scale in 1 study (26); as a reduction in symptoms to no more than mild symptoms during no more than 1 day per week after treatment in 3 studies (24, 25, 29); as a reduction in the use of antacids of more than 75% in 1 study (9); and as control of symptoms without need for additional medications, investigations, or consultations in 1 study (27) (Table 1).

### Patient Characteristics

Data extracted from the studies represent 2793 patients. The average age was similar in all studies. Most included more male than female patients. Symptoms were present for more than 3 months in most patients, and a relatively large proportion of patients had symptoms for longer than 5 years (21, 26, 27). The studies represent the full spectrum of GERD severity, from patients with dyspepsia who have “reflux-like” symptoms to patients who are

Figure 3. Meta-analysis flow diagram.



**Table 3. Diagnostic Evaluation of the Proton-Pump Inhibitor Test with 3 Reference Tests for Gastroesophageal Reflux Disease**

Study (Reference)	Patients					Proportion of Patients with Positive Response	Prevalence of GERD according to Reference Standard	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Positive Likelihood Ratio
	True-Positive Results	False-Negative Results	False-Positive Results	True-Negative Results	Total							
←————— <i>n</i> —————→												
<b>Abnormal 24-hour pH monitoring</b>												
Bate et al. (11)	22	10	11	15	58	0.57	0.55	0.69	0.58	0.67	0.60	1.64
Fass et al. (10)	28	7	3	4	42	0.74	0.83	0.80	0.57	0.90	0.36	1.86
Fass et al. (23)	21	0	8	6	35	0.83	0.60	1.00	0.43	0.72	1.00	1.75
Juul-Hansen et al. (9)	29	5	11	11	56	0.71	0.61	0.85	0.50	0.73	0.69	1.70
Schenk et al. (8)	15	7	7	12	41	0.54	0.54	0.68	0.63	0.68	0.63	1.84
Combined								0.78	0.54			
<b>Esophagitis</b>												
Carlsson et al. (20)	66	72	25	62	225	0.40	0.61	0.48	0.71	0.73	0.46	1.66
Galmiche et al. (24)	27	10	65	39	141	0.66	0.26	0.73	0.38	0.29	0.80	1.18
Hatlebakk et al. (25)	55	22	59	25	161	0.71	0.48	0.71	0.30	0.48	0.53	1.01
Schenk et al. (8)	9	6	13	13	41	0.54	0.37	0.60	0.50	0.41	0.68	1.20
Johnsson et al. (26)	50	17	8	5	80	0.73	0.84	0.74	0.38	0.86	0.23	1.19
Venables et al. (29)	80	21	120	109	330	0.61	0.31	0.79	0.48	0.40	0.84	1.52
Combined								0.71	0.41			

symptomatic but whose symptoms cannot be confirmed by endoscopy to patients with erosive esophagitis (Table 1).

#### Quantitative Data Synthesis

Adequate descriptions of the reference standard and extractable data were provided in 5 studies presenting pH monitoring (8–11, 23) and 6 studies using results of EGD as the primary outcome (8, 20, 24–26, 29). Five studies (19, 21, 22, 27, 28) used a symptom score for selecting patients with GERD (Table 2); however, these studies were not formally combined in the pooled estimates. The symptom scores that were used differed substantially, and no data correlated the various symptom scores to one another, making it unclear whether they can be reasonably combined. The wide range of test characteristics estimated from these studies underscores this concern.

Positive likelihood ratios across studies ranged from 0.45 to 1.87 depending on the reference standard used (Tables 2 and 3). Highest values were observed in studies using abnormal 24-hour pH monitoring as the reference standard or combining this with EGD results. Values were lower when other reference standards were considered; in some studies, the positive likelihood ratio was close to 1.0, suggesting that patients with a symptomatic response to a PPI were just as or even less likely to have GERD compared with those who did not respond. This was especially true in the studies using a symptom score as the reference index. These findings suggest that, at best, the PPI test had only modest discriminant capability, even when considering the most widely accepted reference standard, 24-hour pH monitoring.

Ten studies provided sufficient data for the SROC analysis; one of these studies provided both reference standards (8). An SROC curve of 5 studies using 24-hour pH monitoring as the reference standard (Figure 1) demon-

strates that these studies operate at a slightly higher sensitivity level than the 6 studies using esophagitis in EGD as the reference standard (Figure 2).

The pooled sensitivity and specificity of a positive PPI test result (defined by abnormal 24-hour pH monitoring) were 0.78 (CI, 0.66 to 0.86) and 0.54 (CI, 0.44 to 0.65), respectively (Table 3 and Figure 1). These values were 0.68 (CI, 0.56 to 0.79) and 0.46 (CI, 0.34 to 0.59), respectively, when esophagitis found during EGD was used as the reference standard (Table 3 and Figure 2).

#### DISCUSSION

Although 38% to 90% of the patients suspected of having uncomplicated GERD responded to an empirical trial of treatment with a PPI, our findings suggest that a favorable response does not confidently establish the diagnosis of GERD when GERD is defined according to traditional objective criteria (45). In addition to signaling the presence of GERD, the observed clinical benefit may also indicate the presence of another acid-related condition, a placebo effect, or enhanced esophageal sensitivity to acid exposure (46). On the other hand, 20% to 40% of the patients who have GERD on the basis of objective testing may not exhibit a response to a short course of treatment with a PPI, possibly because they need a higher dose or longer duration of treatment.

Testing for GERD with empirical treatment with a PPI demonstrates only a weak correlation with objective measures. The weakness of the association may in part reflect the limitations of current reference standards (such as pH criteria) for diagnosing GERD (45). Some patients may have symptoms without having pathologic esophageal acid exposure, while others may have abnormal pH results

without having symptoms. The addition of symptom-based or endoscopy-based criteria to the pH results increased sensitivity slightly but did not influence specificity (10, 26).

We also found that successful treatment with a PPI was more likely in patients who had visible erosions (esophagitis) demonstrated on endoscopy, suggesting that patients with GERD complicated by erosive esophagitis represent a population in which the PPI test is more accurate. This finding is consistent with a previous study of short-term treatment of GERD (5). However, this observation has no clinical significance because such patients cannot be identified reliably on the basis of symptoms alone (38).

A striking finding was the poor correlation between treatment success with a PPI and the structured, symptom-based criteria that have been proposed as useful methods for discriminating patients with GERD from those with other causes of symptoms (38, 47). This suggests either that better instruments are needed or that GERD (as defined by symptoms) has substantial overlap with other acid-related disorders, especially in heterogeneous dyspeptic populations (27, 38).

We found that the test characteristics of the PPI test based on endoscopy (as well as on questionnaires) varied more than those of pH-monitoring studies (Table 3). A likely explanation is that interpretation of results of 24-hour pH monitoring is relatively less subjective. However, considerable variability was also observed with studies using 24-hour pH monitoring. The highest sensitivity was observed with studies that used a higher dose of a PPI or with a relatively weak definition of treatment success (9, 23). Lower sensitivities were found in studies with lower doses of a PPI and strong definitions of treatment success (8, 11). Among the studies that used EGD as the reference standard, higher sensitivity was found in studies that defined more severe endoscopic findings as abnormal (24, 29), whereas relatively lower sensitivity was found in studies that considered all endoscopic abnormalities as GERD (20). However, studies with higher sensitivity always had a relatively lower specificity.

A limitation in applying our results to clinical practice is that in reality many patients are given standard doses of a PPI and then monitored for a response after several weeks or even months. In contrast, at least half of the studies included in this meta-analysis used relatively high doses of a PPI and assessed outcomes within 2 weeks. Thus, accuracy of using empirical PPI treatment as a diagnostic test for GERD may be different when lower doses of PPIs are being used and when outcomes are assessed at later time points. It is unlikely, however, that the accuracy would be any better under these conditions because clinical improvement with PPIs is usually observed within a few days and standard doses of PPIs are sufficient to achieve a clinical response in most patients (5, 6, 12). Furthermore, the overall results of our analysis did not change appreciably when we restricted the analysis to studies that used the highest doses of PPIs or when we examined a subset of

studies that assessed outcomes at time points ranging from 2 to 8 weeks (data not shown).

Another limitation of our study is that we had to exclude many trials of PPIs in GERD because the data were presented with insufficient detail for analysis. In particular, few studies clearly described treatment success or the effect of treatment in subgroups. Most treatment studies presented data on the average reduction in specific symptoms but did not provide results on responses in patients without the specific symptoms. Nevertheless, we believe that the relatively few studies that could be included are representative of the types of patients suspected of having GERD in clinical practice.

Review of some of the larger studies not included in our analysis raises questions about the interpretation of study results, even in studies claiming the best test characteristics. For example, a recent study (which reported insufficient detail for inclusion in the quantitative analysis) was carried out with 40 mg of esomeprazole and reported a sensitivity of 83% and a specificity of 30% on day 12 compared with 24-hour pH monitoring as the reference standard (34). The authors concluded that the test performance was sufficient for diagnosing GERD. However, their results translate to a positive likelihood ratio of only 1.19, raising concern about this interpretation.

In conclusion, short-term treatment with a PPI in patients suspected of having GERD does not confidently establish or exclude the diagnosis when GERD is defined by currently accepted reference standards. Given these results, is it reasonable to offer a trial of a PPI in patients suspected of having GERD? Although there may be diagnostic uncertainty, many patients will respond to an empirical trial of a PPI, suggesting that a PPI trial might be reasonable in patients without alarm symptoms or other suspected complications of GERD (3, 4).

On the other hand, the decision to begin with a PPI has long-term economic and clinical implications because responding patients will probably continue treatment even though a diagnosis has not been clearly established. Until better methods are available to establish a confident diagnosis, the empirical treatment approach (and selection of the dose and type of acid-suppressing agents) should be individualized on the basis of the clinical setting, the response to therapy, and judicious diagnostic testing.

From University Medical Center Utrecht, Utrecht, the Netherlands, and Tufts–New England Medical Center, Boston, Massachusetts.

**Acknowledgment:** This work is part of the GERD-guidelines project of the European Society of Primary Care Gastroenterology (ESPCG).

**Grant Support:** Mattijs Numans, MD, PhD, was financially supported for this study in the United States by the Netherlands Organization for Scientific Research, the University Medical Centre, Utrecht, the Netherlands (Foundation Girard de Miele van Coehoorn), and by unrestricted grants from the Dutch divisions of Altana Pharma, Jansen Cilag,

and AstraZeneca. Joseph Lau, MD, is supported in part by a grant (R01 HS013328) from the Agency for Healthcare Research and Quality.

**Potential Financial Conflicts of Interest:** *Stock ownership or options (other than mutual funds):* J. Lau (Merck, Pfizer); *Expert testimony:* N.J. de Wit (Janssen Cilag); *Grants received:* M.E. Numans (AstraZeneca, Altana Pharma, Janssen Cilag), N.J. de Wit (Altana Pharma, Novartis, Janssen Cilag).

**Requests for Single Reprints:** Mattijs E. Numans, MD, PhD, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85060, 3508 AB Utrecht, the Netherlands; e-mail, m.e.numans@med.uu.nl.

Current author addresses and author contributions are available at [www.annals.org](http://www.annals.org).

## References

- Locke GR 3rd, Talley NJ, Fett SL, Zinsmeister AR, Melton LJ 3rd. Prevalence and clinical spectrum of gastroesophageal reflux: a population-based study in Olmsted County, Minnesota. *Gastroenterology*. 1997;112:1448-56. [PMID: 9136821]
- van Bommel MJ, Numans ME, de Wit NJ, Stalman WA. Consultations and referrals for dyspepsia in general practice—a one year database survey. *Postgrad Med J*. 2001;77:514-8. [PMID: 11470932]
- DeVault KR, Castell DO. Updated guidelines for the diagnosis and treatment of gastroesophageal reflux disease. The Practice Parameters Committee of the American College of Gastroenterology. *Am J Gastroenterol*. 1999;94:1434-42. [PMID: 10364004]
- Kroes RM, Numans ME, Jones RH, de Wit NJ, Verheij TJ. Gastro-oesophageal reflux disease in primary care. Comparison and evaluation of existing national guidelines and development of uniform European guidelines. *European Journal of General Practise*. 1999;5:88-97.
- van Pinxteren B, Numans ME, Lau J, de Wit NJ, Hungin AP, Bonis PA. Short-term treatment of gastroesophageal reflux disease. *J Gen Intern Med*. 2003;18:755-63. [PMID: 12950485]
- Fass R. Empirical trials in treatment of gastroesophageal reflux disease. *Dig Dis*. 2000;18:20-6. [PMID: 10729734]
- Schindlbeck NE, Klauser AG, Voderholzer WA, Muller-Lissner SA. Empiric therapy for gastroesophageal reflux disease. *Arch Intern Med*. 1995;155:1808-12. [PMID: 7654116]
- Schenk BE, Kuipers EJ, Klinkenberg-Knol EC, Festen HP, Jansen EH, Tuynman HA, et al. Omeprazole as a diagnostic tool in gastroesophageal reflux disease. *Am J Gastroenterol*. 1997;92:1997-2000. [PMID: 9362179]
- Juul-Hansen P, Rydning A, Jacobsen CD, Hansen T. High-dose proton-pump inhibitors as a diagnostic test of gastro-oesophageal reflux disease in endoscopic-negative patients. *Scand J Gastroenterol*. 2001;36:806-10. [PMID: 11495074]
- Fass R, Ofman JJ, Gralnek IM, Johnson C, Camargo E, Sampliner RE, et al. Clinical and economic assessment of the omeprazole test in patients with symptoms suggestive of gastroesophageal reflux disease. *Arch Intern Med*. 1999;159:2161-8. [PMID: 10527293]
- Bate CM, Riley SA, Chapman RW, Durnin AT, Taylor MD. Evaluation of omeprazole as a cost-effective diagnostic test for gastro-oesophageal reflux disease. *Aliment Pharmacol Ther*. 1999;13:59-66. [PMID: 9892880]
- An evidence-based appraisal of reflux disease management—the Genval Workshop Report. *Gut*. 1999;44 Suppl 2:S1-16. [PMID: 10741335]
- Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol*. 2000;53:65-9. [PMID: 10693905]
- Ioannidis JP, Chew P, Lau J. Standardized retrieval of side effects data for meta-analysis of safety outcomes. A feasibility study in acute sinusitis. *J Clin Epidemiol*. 2002;55:619-26. [PMID: 12063104]
- Numans ME, de Wit NJ. Reflux symptoms in general practice: diagnostic evaluation of the Carlsson-Dent gastro-oesophageal reflux disease questionnaire. *Aliment Pharmacol Ther*. 2003;17:1049-55. [PMID: 12694087]
- Numans ME, Van der Graaf Y, de Wit NJ, Touw-Otten F, de Melker RA. How much ulcer is ulcer-like? Diagnostic determinants of peptic ulcer in open access gastroscopy. *Fam Pract*. 1994;11:382-8. [PMID: 7895965]
- Demeester TR, Johnson LF, Joseph GJ, Toscano MS, Hall AW, Skinner DB. Patterns of gastroesophageal reflux in health and disease. *Ann Surg*. 1976;184:459-70. [PMID: 13747]
- Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med*. 1989;4:288-95. [PMID: 2760697]
- Brun J, Sorngard H. High dose proton pump inhibitor response as an initial strategy for a clinical diagnosis of gastro-oesophageal reflux disease (GERD). Swedish multi-centre group in primary health care. *Fam Pract*. 2000;17:401-4. [PMID: 11021899]
- Carlsson R, Dent J, Watts R, Riley S, Sheikh R, Hatlebakk J, et al. Gastro-oesophageal reflux disease in primary care: an international study of different treatment strategies with omeprazole. International GORD Study Group. *Eur J Gastroenterol Hepatol*. 1998;10:119-24. [PMID: 9581986]
- Dupas JL, Houcke P, Samoyeau R, French Collaborative Pantoprazole Study Group. Pantoprazole versus lansoprazole in French patients with reflux esophagitis. *Gastroenterol Clin Biol*. 2001;25:245-50. [PMID: 11395670]
- Farup PG, Hovde O, Torp R, Wetterhus S. Patients with functional dyspepsia responding to omeprazole have a characteristic gastro-oesophageal reflux pattern. *Scand J Gastroenterol*. 1999;34:575-9. [PMID: 10440606]
- Fass R, Ofman JJ, Sampliner RE, Camargo L, Wendel C, Fennerty MB. The omeprazole test is as sensitive as 24-h oesophageal pH monitoring in diagnosing gastro-oesophageal reflux disease in symptomatic patients with erosive oesophagitis. *Aliment Pharmacol Ther*. 2000;14:389-96. [PMID: 10759617]
- Galmiche JP, Barthelemy P, Hamelin B. Treating the symptoms of gastro-oesophageal reflux disease: a double-blind comparison of omeprazole and cispripide. *Aliment Pharmacol Ther*. 1997;11:765-73. [PMID: 9305487]
- Hatlebakk JG, Hyggen A, Madsen PH, Walle PO, Schulz T, Mowinckel P, et al. Heartburn treatment in primary care: randomised, double blind study for 8 weeks. *BMJ*. 1999;319:550-3. [PMID: 10463897]
- Johnsson F, Weywadt L, Solhaug JH, Hernqvist H, Bengtsson L. One-week omeprazole treatment in the diagnosis of gastro-oesophageal reflux disease. *Scand J Gastroenterol*. 1998;33:15-20. [PMID: 9489902]
- Lewin van den Broek NT, Numans ME, Buskens E, Verheij TJ, de Wit NJ, Smout AJ. A randomised controlled trial of four management strategies for dyspepsia: relationships between symptom subgroups and strategy outcome. *Br J Gen Pract*. 2001;51:619-24. [PMID: 11510389]
- Talley NJ, Meineche-Schmidt V, Pare P, Duckworth M, Raisanen P, Pap A, et al. Efficacy of omeprazole in functional dyspepsia: double-blind, randomized, placebo-controlled trials (the Bond and Opera studies). *Aliment Pharmacol Ther*. 1998;12:1055-65. [PMID: 9845395]
- Venables TL, Newland RD, Patel AC, Hole J, Wilcock C, Turbitt ML. Omeprazole 10 milligrams once daily, omeprazole 20 milligrams once daily, or ranitidine 150 milligrams twice daily, evaluated as initial therapy for the relief of symptoms of gastro-oesophageal reflux disease in general practice. *Scand J Gastroenterol*. 1997;32:965-73. [PMID: 9361167]
- Hatlebakk JG, Berstad A, Carling L, Svedberg LE, Unge P, Ekstrom P, et al. Lansoprazole versus omeprazole in short-term treatment of reflux oesophagitis. Results of a Scandinavian multicentre trial. *Scand J Gastroenterol*. 1993;28:224-8. [PMID: 8446846]
- Savary M, Miller G. The Oesophagus. Solothurn, Switzerland: Grassmann; 1977.
- Kahrilas PJ, Quigley EM. Clinical esophageal pH recording: a technical review for practice guideline development. *Gastroenterology*. 1996;110:1982-96. [PMID: 8964428]
- Weusten BL, Roelofs JM, Akkermans LM, Van Berge-Henegouwen GP, Smout AJ. The symptom-association probability: an improved method for symptom analysis of 24-hour esophageal pH data. *Gastroenterology*. 1994;107:1741-5. [PMID: 7958686]
- Johnsson F, Hatlebakk JG, Klintonberg AC, Roman J, Toth E, Stubberöd A, et al. One-week esomeprazole treatment: an effective confirmatory test in patients with suspected gastroesophageal reflux disease. *Scand J Gastroenterol*. 2003;38:354-9. [PMID: 12739706]
- Hetzel DJ, Dent J, Reed WD, Narielvala FM, Mackinnon M, McCarthy JH, et al. Healing and relapse of severe peptic esophagitis after treatment with omeprazole. *Gastroenterology*. 1988;95:903-12. [PMID: 3044912]

36. **Ollyo JB, Lang F, Fontollet C, Monnier P.** Savary-Miller's new reproducible, logical, complete and useful classification [Abstract]. *Gastroenterology*. 1990;98:A100.
37. **Orlando RC.** Esophageal epithelial defenses against acid injury. *Am J Gastroenterol*. 1994;89(8 Suppl):S48-52. [PMID: 8048414]
38. **Bytzer P, Hansen JM, Schaffalitzky de Muckadell OB, Malchow-Moller A.** Predicting endoscopic diagnosis in the dyspeptic patient. The value of predictive score models. *Scand J Gastroenterol*. 1997;32:118-25. [PMID: 9051871]
39. **Moses LE, Shapiro D, Littenberg B.** Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*. 1993;12:1293-316. [PMID: 8210827]
40. **Littenberg B, Moses LE.** Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making*. 1993;13:313-21. [PMID: 8246704]
41. **Swets JA.** Measuring the accuracy of diagnostic systems. *Science*. 1988;240:1285-93. [PMID: 3287615]
42. **Shapiro DE.** Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol*. 1995;2 Suppl 1:S37-47; discussion S65-9, S83. [PMID: 9419704]
43. **Lau J.** Meta-Test, version 0.6. Boston: New England Medical Center; 1997.
44. **Bonis PA, Ioannidis JP, Cappelleri JC, Kaplan MM, Lau J.** Correlation of biochemical response to interferon alfa with histological improvement in hepatitis C: a meta-analysis of diagnostic test characteristics. *Hepatology*. 1997;26:1035-44. [PMID: 9328332]
45. **van Herwaarden MA, Smout AJ.** Diagnosis of reflux disease. *Baillieres Best Pract Res Clin Gastroenterol*. 2000;14:759-74. [PMID: 11003808]
46. **Watson RG, Tham TC, Johnston BT, McDougall NI.** Double blind cross-over placebo controlled study of omeprazole in the treatment of patients with reflux symptoms and physiological levels of acid reflux—the "sensitive oesophagus". *Gut*. 1997;40:587-90. [PMID: 9203934]
47. **Klauser AG, Schindlbeck NE, Muller-Lissner SA.** Symptoms in gastro-oesophageal reflux disease. *Lancet*. 1990;335:205-8. [PMID: 1967675]

---

**Current Author Addresses:** Drs. Numans and de Wit: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85060, 3508 AB Utrecht, the Netherlands.

Drs. Lau and Bonis: Center for Clinical Evidence Synthesis, Tufts–New England Medical Center, Kneeland Drive, Boston MA 02111.

**Author Contributions:** Conception and design: M.E. Numans, J. Lau, N.J. de Wit, P.A. Bonis.

Analysis and interpretation of the data: M.E. Numans, J. Lau, P.A. Bonis.

Drafting of the article: M.E. Numans, P.A. Bonis.

Critical revision of the article for important intellectual content: M.E. Numans, J. Lau, N.J. de Wit, P.A. Bonis.

Final approval of the article: M.E. Numans, J. Lau, P.A. Bonis.

Provision of study materials or patients: M.E. Numans.

Statistical expertise: M.E. Numans, J. Lau, P.A. Bonis.

Obtaining of funding: M.E. Numans, N.J. de Wit.

Administrative, technical, or logistic support: J. Lau.

Collection and assembly of data: M.E. Numans, P.A. Bonis.