THE AMERICAN JOURNAL OF GASTROENTEROLOGY
Vol. 96, No. 1, 2001

ISSN 0002-9270/01/$20.00
Published by Elsevier Science Inc.
PII S0002-9270(00)02244-9

Initial Validation of a Diagnostic Questionnaire for Gastroesophageal Reflux Disease

Michael J. Shaw, M.D., F.A.C.G., Nicholas J. Talley, M.D., Ph.D., F.A.C.G., Timothy J. Beebe, Ph.D., Todd Rockwood, Ph.D., Rolf Carlsson, Ph.D., Susan Adlis, M.S., A. Mark Fendrick, M.D., Roger Jones, M.D., John Dent, M.D., Ph.D., and Peter Bytzer, M.D., Ph.D.

Health Research Center, Park Nicollet Medical Center, Minneapolis, Minnesota; Department of Medicine, The University of Sydney, The Nepean Hospital, Penrith, Australia; Clinical Outcomes Research Center, University of Minnesota, Minneapolis, Minnesota; AstraZeneca, Mölndal, Sweden; Department of Medicine, University of Michigan Medical School, Ann Arbor, Michigan; Division of Primary Care and Public Health Services, Guy's, King's, and St. Thomas' School of Medicine, London, UK; Department of Gastrointestinal Medicine, Royal Adelaide Hospital, North Terrace Adelaide, South Australia; and Department of Medical Gastroenterology, Glostrup University Hospital, Glostrup, Denmark

OBJECTIVES: Brief, reliable, and valid self-administered questionnaires could facilitate the diagnosis of gastroesophageal reflux disease in primary care. We report the development and validation of such an instrument.

METHODS: Content validity was informed by literature review, expert opinion, and cognitive interviewing of 50 patients resulting in a 22-item survey. For psychometric analyses, primary care patients completed the new questionnaire at enrollment and at intervals ranging from 3 days to 3 wk. Multitrait scaling, test–retest reliability, and responsiveness were assessed. Predictive validity analyses of all scales and items used specialty physician diagnosis as the "gold standard."

RESULTS: Iterative factor analyses yielded three scales of four items each including heartburn, acid regurgitation, and dyspepsia. Multitrait scaling criteria including internal consistency, item interval consistency, and item discrimination were 100% satisfied. Test–retest reliability was high in those reporting stable symptoms. Scale scores significantly changed in those reporting a global change. Regressing specialty physician diagnosis on the three scales revealed significant effects for two scales (heartburn and regurgitation). Combining the two significant scales enhanced the strength of the model. Symptom response to self-directed treatment with nonprescription antisecretory medications was highly predictive of the diagnosis also, although the item demonstrated poor validity and reliability.

CONCLUSIONS: A brief, simple 12-item questionnaire demonstrated validity and reliability and seemed to be responsive to change for reflux and dyspeptic symptoms. (Am J Gastroenterol 2001;96:52–57. © 2001 by Am. Coll. of Gastroenterology)

INTRODUCTION

Recent research on diagnostic methodology for gastroesophageal reflux disease (GERD) has provided two important observations. First, most individuals seen by primary care physicians for GERD do not have significant esophagitis (1). Second, it has been recognized that there is not a criterion standard for diagnosis of the illness (2, 3). Although initially proposed as a "gold standard" for the diagnosis of GERD, 24-h ambulatory pH monitoring is not sufficiently sensitive to serve as a criterion standard (2, 4). Recent evidence has suggested that a brief treatment trial with a proton pump inhibitor may be the most sensitive and specific diagnostic modality (2, 5). However, studies of that potential diagnostic test have some limitations, including the fact that the survey instrument supporting the test has not been validated for the diagnosis of GERD.

A number of surveys that include items pertinent to GERD have been developed and published in the past 5 yr (6–11). Five of these surveys are of established reliability and validity for assessment of symptom severity (6–10) and health-related quality of life (6–9). Three have been shown to be sensitive to change with treatment of GERD (7, 8, 12). None has been evaluated for discriminant (diagnostic) validity for GERD, although all would be good candidates for such studies.

Discriminant validity has been briefly examined in one of these surveys (11). Cognitive interviewing or focus group testing of patients was not included in the initial design of that instrument, and the study population was highly enriched for reflux disease. Scale scores were not generated with simple addition, but required differential weighting of various items. Empiric support for the selected item weights was not provided. An important contribution of that study was recognition that the word "heartburn" is less sensitive

than a word picture of the symptom ("a burning feeling rising from your stomach or lower chest up toward your neck") (11).

The Digestive Health Status Instrument (DHSI) GERD/Ulcer scale has been compared to 24-h ambulatory pH monitoring in a population of referral patients, and the scale scores correlated highly with acid contact time (12). The DHSI is relatively long at 34 questions, limiting routine application in a primary care setting.

We report the development and validation of a brief instrument for the diagnosis of GERD that is simple to administer, to score, and to interpret.

MATERIALS AND METHODS

Setting

From July 14, 1999 to September 17, 1999, two separate subject groups were enlisted: one to participate in questionnaire development and the other in questionnaire validation. Adults aged ≥18 yr were recruited from various primary care clinics of Park Nicollet Clinic, HealthSystem Minnesota, Minneapolis, MN. Park Nicollet is a 500 physician multispecialty group serving the western and southern suburbs of Minneapolis, MN.

Survey Development

Content validity of the survey was established following the method of Dillman (13). The literature was searched for survey instruments and the pertinent ones were carefully reviewed (6–11). The questionnaire was drafted by the principal investigator and a survey methodologist on the research team. The core elements of this first questionnaire possessed many similarities to the Carlsson-Dent reflux scale (11). The survey content and wording were modified by the other members of the research team. The questions were scaled.

Cognitive interviews on the content and wording of the new survey using a planned script were conducted in 25 patients presenting to discuss reflux symptoms with their primary care physician. Patients received a $25 honorarium for their participation. Patients were interviewed by a survey technician after they had completed both questionnaires. Subjects were queried as to whether the items adequately surveyed the breadth of their symptoms. Potential problems with the design or wording of questions and interpretation were sought.

Results of the cognitive interviewing were presented at a meeting of all coinvestigators. Changes informed by the cognitive interviewing were incorporated into the questionnaire to improve clarity.

The reworded questionnaire was presented to 10 of these 25 subjects in two focus group sessions for review. No additional revisions were suggested. Cognitive interviews as described above were conducted in a different group of 25 patients. No additional changes were suggested.

Survey Validation

For the validation of the new survey, 200 consecutive Park Nicollet Clinic primary care patients from nine separate clinic sites in the southern and western suburbs of Minneapolis, MN, were recruited. All had appointments with their family practice or internal medicine physician to discuss upper abdominal complaints that might include but did not require the presence of reflux symptoms. Eligible patients were identified by a three-item questionnaire at registration asking whether the patient was visiting the healthcare provider to discuss upper abdominal pain or heartburn, whether he or she had taken prescription medication for symptoms in the past month, and general interest in completing a short survey. Those who had received prescription medication in the month preceding entry were excluded. Subjects who completed the survey on two occasions and a specialty physician interview received a $50 honorarium. The survey included the 22-item pilot version of the Reflux Disease Diagnostic Questionnaire (RDQ) and the Digestive Health Status Instrument (DHSI). Each of these measures is described in more detail in the following sections. Surveys were completed and returned. To determine whether they had GERD, subjects were interviewed by a board-certified gastroenterologist. Power calculations revealed a need for 150 subjects. These 150 subjects were randomly selected from the pool of 200 to ensure that there were no differences in age and gender of the smaller subset compared to the complete study population.

MEASURES. The new instrument, in its pilot form, had 22 items. The time referent used was symptoms over the last 4 wk. Item content included the following: 1) 12 questions on the frequency, severity, and duration of burning and pain behind breastbone, acid taste in mouth, movement of materials upward from the stomach, and burning and pain in the upper stomach; 2) one question asking for the location where symptoms were most severe; 3) two questions on whether symptoms were caused or worsened by eating, lying flat, or bending over; 4) two questions on relief of symptoms by indigestion medications or eating; 5) three questions on the interference of symptoms in patients' sleep, work, and enjoyment of life; and 6) one question asking which symptom caused the greatest interference with life. Most response options were scaled as closed-ended and Likert-type, with categories ranging from 1 to 5 or with 1–7 points for frequency, severity, and duration of symptoms. The item assessing the area affected contained a diagram of a human torso, and the patient was asked to mark the area most affected.

The DHSI is a 34-item, disease-specific instrument with demonstrated reliability and validity (6). Five separate constructs measured by the DHSI include GERD/Ulcer, dysmotility, a two-domain bowel dysfunction complex, and a pain index. The psychometric properties of these five summated disease-specific scales compared favorably with standardized health status measures. As was the case with the

RDQ, the time referent of the DHSI is symptom occurrence over the last 4 wk and response scales are Likert-type and closed-ended with values ranging usually from 1 to 6.

RESPONSIVENESS QUESTIONS. To assess the responsiveness of the RDQ, patients were asked to rate their upper abdominal symptoms since their last visit on a three-point scale including better, unchanged, and worse. If their symptoms had changed, patients were asked to rate how much they had changed on a seven-point scale ranging from changed hardly at all to changed a great deal. This has been referred to as the observed treatment effect (OTE) (14). The patient's physician was unaware of participation in the study, and patient management was provided at the discretion of the individual physician. Many patients received antiscretory therapy at the enrollment visit.

SPECIALTY PHYSICIAN DIAGNOSIS. A random selection of 75% of the study population was interviewed by a gastroenterologist to determine if GERD was present. Six board-certified gastroenterologists participated in the interviewing. The physicians were aware that the study objective was to develop a brief reflux disease survey but did not have knowledge of the questionnaire content or wording, the study protocol, or the individual subject responses to the questionnaire. Interrater reliability of the group of physicians was determined before the subjects were interviewed. Ten cases of established diagnoses including GERD alone (4), irritable bowel syndrome (IBS) alone (3), and GERD-IBS (3) were identified from the records of Park Nicollet Clinic. Five actors were trained to present two cases each and were interviewed individually by the six physicians. Physicians were asked to decide whether the patient had GERD, IBS, or neither and to provide certainty of the diagnosis. Agreement on diagnosis from interviews of the 10 cases was determined and κ calculated (15). The κ levels reached acceptable levels for the diagnosis of GERD (0.90) and IBS (0.80).

Psychometric Evaluation

There were three primary objectives in the psychometric evaluation of the RDQ: 1) item reduction and scale development; 2) convergent and predictive validity assessment; and 3) stability and responsiveness assessment. Item reduction and scale development employed approaches associated with multitrait scaling, including item–response variability, factor analytic testing, scale internal consistency, item convergent validity, item discrimination validity, and item–total correlations (16). Convergent validity assessment looked at correlations between RDQ scales and those contained in the DHSI. Similar comparisons of the RDQ scales with specialty physician diagnosis were conducted to assess the predictive validity of the RDQ. The predictive capability of individual questions was also assessed. The stability of the RDQ scales was examined by test–retest reliability through construction of intraclass correlation coefficients in those reporting no change in symptoms, and responsiveness was determined in those reporting at least moderate change using the *t* statistic for paired samples. Assignment to an unchanged or improved group was based on response to the observed treatment effect question (14).

RESULTS

For the instrument validation, a total of 200 patients meeting the inclusion criteria were enrolled in the study. Of these, 176 (88%) patients filled out the initial packet of surveys that included the RDQ and the DHSI. The participant pool included 123 (70%) women and 53 (30%) men. The mean age of responders was 50.5 yr. One respondent was an African-American woman and another an Asian man; all other study participants were Caucasian. A total of 150 of these respondents were interviewed by a specialty physician; 92% of those interviewed completed the questionnaire a second time at intervals of 3 days to 3 wk. Physician diagnoses were GERD in 69%, IBS in 37%, GERD and IBS in 17%, and neither in 17%.

Content Validity

Additional content was not suggested by any of the 50 persons in the questionnaire development cohort; however, consistent themes emerged from the cognitive interviewing, indicating a need for significant rewording of the questions. The first was the challenge for patients in responding to questions requiring integration of more than two pieces of information in one question, *e.g.*, asking for the quality ("burning"), location ("stomach or lower chest"), and movement of a symptom ("rising") in one question. The second concerned describing reflux symptoms as "rising upward." Only 20% of those interviewed with GERD reported upward movement, and patients were as likely to report downward or a posterior movement to their reflux symptoms. The percentage of subjects without GERD reporting upward movement of their symptoms nearly equaled that found in those with GERD (15% *vs* 20%, respectively). Selection of location was problematic; 80% specified that separating symptom location in the upper stomach from lower chest was preferred over combining into one question. These observations were incorporated into the pilot questionnaire after review by the entire group of coinvestigators.

Item Reduction and Scale Development

ITEM REDUCTION. A total of 10 items were deleted from the original 22-item pilot RDQ. Two items measuring relief of symptoms by eating and symptoms that caused the greatest interference with life were discarded because they failed to meet the item-response criterion requiring that response distributions and standard deviations be roughly symmetrical in items measuring the same construct. Four more items were deleted because they failed to meet the factor analytic interpretability criteria requiring item loadings of >0.4 on only one component for an item to be retained. The items included areas most affected by symptoms, symptoms

Table 1. Results of the Principal Component Analysis and Internal Consistency Reliability Analysis

BreDQ Question	Regurgitation	GERD	Dyspepsia
2c. Acid taste severity	0.89796	−0.10826	−0.02118
2d. Movement of materials severity	0.83424	0.00360	−0.00187
1c. Acid taste frequency	0.79231	0.00545	0.07755
1d. Movement of materials frequency	0.77503	0.03962	0.02054
1a. Frequency of burning behind breastbone	0.02244	0.88609	−0.15100
2a. Severity of burning behind breastbone	0.12303	0.87884	−0.18781
1b. Frequency of pain behind breastbone	−0.14538	0.71539	0.34627
2b. Severity of pain behind breastbone	−0.11166	0.65196	0.27500
2f. Upper stomach pain severity	−0.02989	−0.10742	0.93534
1f. Upper stomach pain frequency	−0.01706	−0.04117	0.92785
1e. Upper stomach burning frequency	0.26315	0.24351	0.47527
2e. Upper stomach burning severity	0.32110	0.16140	0.45311
Cronbach's α	0.85	0.81	0.80

GERD = gastroesophageal reflux disease.

caused or made worse by eating, relief of symptoms through use of indigestion medications, and symptoms caused or made worse by lying flat or bending over. Finally, the item measuring duration of symptoms and the three items measuring symptom interference with sleep, work, and life enjoyment (quality of life) were discarded because of their questionable ability to forecast specialty physician diagnosis at the item level (the scale–level predictive validity assessment described in detail below).

SCALE DEVELOPMENT. To examine the dimensionality of the remaining 12 items, principal component analysis was repeated. From one to four component models were considered. A three-factor solution was considered most plausible and is shown in Table 1.

The first component contains four items that were thought to measure the domain of "regurgitation." Constituent items address the frequency and severity of acid taste and movement of materials. The second four-item component, labeled "heartburn," is composed of items measuring frequency and severity of burning and pain behind the breastbone. The final component measures dyspeptic symptoms. Constituent items include upper stomach pain severity, upper stomach pain frequency, upper stomach burning frequency, and upper stomach burning severity.

Cronbach's α was calculated to demonstrate the cohesiveness of items in each extracted component (17). All three components demonstrated high internal consistency, with α scores exceeding the acceptable level of 0.70 (see Table 1) (18).

Item–scale correlations were examined in a matrix in which the items are in rows and the scales are in columns. Correlations between items and scales were corrected for overlap. Analyses of these correlations, which are themselves a measure of item convergent validity, showed that each item had a correlation with its relevant scale of $\geq$0.40 (data not shown). Similarly, the item discrimination criterion (discriminant validity) requiring that the correlation between an item and its hypothesized scale be greater than two standard errors (SE) larger than any other correlation in the same row to consider it a scaling success was also satisfied (data not shown).

The last component of multitrait scaling, the assessment of item–total correlations, requires that items in the same scale contain the same proportion of information about the construct. The range of correlations corrected for overlap for each scale was examined. Items identified in the principal component analysis to belong to a scale correlated most closely with the hypothesized scale for every single item. Regurgitation correlations ranged from 0.66–0.76, heartburn 0.61–0.70, and dyspepsia 0.58–0.65.

Multitrait scaling criteria were 100% satisfied by the three scales, justifying the simple addition of subject responses to obtain scale scores.

Convergent and Predictive Validity Assessment

This step attempted to assess the convergent validity of the RDQ by looking at how well the identified scales of the RDQ correlated with another scale that is believed to measure the same construct: the GERD/Ulcer scale of the DHSI. Evidence for the predictive validity of the RDQ (meaning how well it forecasts specialty physician diagnosis) was also sought in this step.

CONVERGENT VALIDITY. The extent to which the identified scales correlated with the DHSI scales was assessed by calculating Pearson product-moment coefficients. We hypothesized that the GERD/Ulcer scale of the DHSI would correlate highly with the heartburn and regurgitation scales of the new survey. Table 2 shows that the highest correlation was observed for the relationship between the RDQ heartburn and DHSI GERD/Ulcer scales (0.52).

Table 2. Convergent Validity: BreDQ and DHSI Correlations

	BreDQ Domain		
DHSI Domain	Regurgitation	Heartburn	Dyspepsia
Heartburn + ulcer	0.42	0.52	0.30
Pain experience	0.17	0.33	0.47

DHSI = Digestive Health Status Instrument.

Table 3. Predictive Validity: Logistic Regression of BreDQ Scale Scores on Specialty Physician Diagnosis

Scale	Physician Diagnosis: GERD Mean (SD)	Physician Diagnosis: No GERD Mean (SD)	*p* Value*	c Statistic
Heartburn	8.7 (5.2)	6.2 (5.4)	0.01	0.64
Regurgitation	6.6 (5.0)	4.3 (4.1)	0.01	0.64
Dyspepsia	7.1 (5.1)	7.7 (5.3)	0.55	0.53
OTC response	3.2 (2.3)	2.3 (0.9)	0.0076	0.70

* Based on logistic regression analysis.
GERD = gastroesophageal reflux disease; OTC = over-the-counter medication.

PREDICTIVE VALIDITY. Table 3 provides the mean scores for each of the three RDQ scales in those with and without GERD by physician diagnosis; the significance of the scale's ability to predict specialty physician diagnosis of GERD; and the c-statistic, a measure of the scale's ability to discriminate between those meeting the diagnostic threshold and those who do not. Both the heartburn and regurgitation scales significantly predicted physician diagnosis at the $p = 0.01$ level. The dyspepsia scale failed to provide any evidence of predictive power. The c-statistic showed moderate accuracy (0.64) of the heartburn and regurgitation scales in classifying patients into one of the two diagnostic categories.

The strongest predictor of physician diagnosis was response of symptoms to various over-the-counter medications ($p = 0.0076$, with c-statistic of 0.70).

Stability and Responsiveness

TEST–RETEST RELIABILITY. In all, 58 subjects reported no change in symptoms 3 days to 3 wk after entry on the observed treatment effect question (17). Intraclass correlation coefficients were calculated from the questionnaires of those who reported no change in symptoms over the period of observation. All three RDQ scales demonstrated reliability coefficients ranging from 0.8 to 0.88 well beyond the acceptable level of 0.70 (data not shown) (18).

RESPONSIVENESS. Responsiveness was determined in the subset of 59 respondents who indicated at least moderate change on the observed treatment effect. The Student's *t* test for paired samples was calculated comparing those who remained stable to those reporting at least moderate change. The *p* values were ≤ 0.0029 for all three scales, providing strong evidence for the responsiveness of each of the three scales of the RDQ (data not shown).

DISCUSSION

Progress in studying the medical and financial impact of gastroesophageal reflux disease (GERD) is compromised by the lack of patient-report questionnaires validated for the diagnosis of GERD. Methodological (*e.g.*, limited psychometric evaluation) and administrative (*e.g.*, length) shortcomings associated with earlier attempts to develop such instruments limit the potential application of these instruments. The purpose of this investigation was to develop a reliable and valid instrument for the diagnosis of GERD that could be easily administered by primary care physicians in community settings.

Developing a questionnaire with excellent content validity requires careful attention to both question content and clarity of wording. No amount of sophisticated psychometrics can compensate for deficiencies in what—and how—questions are asked. Question selection and wording were informed by a review of the literature, input from a number of investigators who have developed and validated GI symptom questionnaires, and, most importantly, detailed cognitive interviewing of patients. Despite the prior development of many reflux questionnaires (6–12), cognitive interviewing demonstrated that the wording of some questions in these questionnaires created interpretation difficulties for patients, highlighting the need for focus group testing or cognitive interviewing during the development of any new surveys.

Scale development provided empirical support for three distinct symptom profiles: heartburn, regurgitation, and dyspepsia. Item reduction efforts nearly halved the number of items in the survey. Scale reliability was demonstrated by the high levels of internal consistency. Strong support for other elements of multitrait scaling including item–scale correlations, item discrimination, and item–total correlations was observed. Complete satisfaction of multitrait scaling criteria justifies combining the items into scales that can be scored with simple addition, thus eliminating the need for item weighting.

The high correlations between the GERD/Ulcer scale of the Digestive Health Status Instrument (DHSI) and the heartburn scale of the RDQ attest to the convergent validity of the RDQ. The ability of the heartburn and regurgitation scales to predict specialty physician diagnosis demonstrates the predictive validity of at least two of the three scales of the RDQ; the dyspepsia scale failed to demonstrate acceptable predictive validity. The lack of predictive validity of the dyspepsia scale and the principal component analyses demonstrating that symptoms in the upper stomach are a distinct construct from those located in the retrosternal area validates the results of the cognitive interviewing, leading to a decision to include symptom locations.

The dyspepsia scale demonstrated excellent internal validity and was responsive to change. Its psychometric strength combined with information that primary care and community populations frequently demonstrate overlapping symptom complexes, including elements of reflux and dyspepsia (19), support retaining the dyspepsia scale pending future research using this instrument. The scales proposed for GERD diagnosis, however, are limited to the regurgitation and heartburn scales.

A gold standard for the diagnosis of GERD does not exist. The sensitivity of endoscopy is limited, as most patients with GERD do not have mucosal injury (1). Ambulatory pH monitoring also has problems with sensitivity, given the intermittent nature of symptoms and the disturbance of routine daily activities by placement of a pH probe (2–4). Specialty physician diagnosis is an acceptable gold standard in the absence of a definitive objective test (20, 21). The six

gastroenterologists demonstrated excellent agreement, both among themselves and on patient diagnosis, when interviewing actors presenting cases of established diagnoses, thus supporting the use of specialty physician diagnosis as a gold standard. High correlations between the GERD/Ulcer scale of the DHSI and the RDQ provide indirect support for the validity of this approach given the established correlation of the GERD/Ulcer scale of the DHSI with ambulatory pH monitoring in GERD patients (12).

One intriguing finding was that the strongest predictor of specialty physician diagnosis was the question on the response of symptoms to self-directed, over-the-counter medications. The predictive power of this single question surpassed that of any of the three scales alone, any combination of the three scales, or the other 20 individual items in the questionnaire (data not shown). Despite this, however, the question demonstrated poor internal validity as well as internal consistency below an acceptable level, resulting in its deletion from the final survey. The work of Fass *et al.* and of others suggests that a more structured approach to assessing response maybe the preferred diagnostic method for GERD (2, 4). The studies of diagnostic treatment trials have possessed a number of methodological shortcomings, however, including the lack of valid and reliable surveys to assess treatment effect. A diagnostic trial using the RDQ to assess response to standardized treatment with superior acid suppression would retain the observed predictive strength of the response to OTC medications, but without compromising validity and reliability.

Although these results are encouraging, readers should be mindful of some study limitations. The stability and responsiveness of the instrument was examined in a minority of the research subjects for whom treatment was directed by individual physicians and not by study design. Generalization of results from a study performed in only one location among a Caucasian population to other areas with different racial/ethinic populations cannot be assumed. Although there is precedent for using specialty physician diagnosis as a gold standard for disease diagnosis, this approach has not been definitively validated for GERD. Studies addressing these limitations are underway, including the use of questionnaires in other countries besides the United States and in other languages in addition to English.

A brief survey for the diagnosis of GERD in primary care and community settings that can be simply administered and scored has been developed and validated. The identified predictive strength of response to treatment argues in support of a diagnostic treatment trial. The discriminative (diagnostic) validity of the heartburn and regurgitation scales will be definitively established in such a study.

Reprint requests and correspondence: Michael J. Shaw, M.D., Health Research Center, Park Nicollet Medical Center, 3800 Park Nicollet Boulevard, Minneapolis, MN 55416-2699.

Received June 12, 2000; accepted Sep. 14, 2000.

REFERENCES

1. Jones RH, Hungin PAS, Phillips J, et al. Gastro-oesophageal reflux disease in primary care in Europe: Clinical presentation and endoscopic findings. Eur J Gen Pract 1995;1:149–54.
2. Fass R, Ofman JJ, Gralnek IM, 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.
3. Dent J, Brun J, Fendrick AM, et al. An evidence-based appraisal of reflux disease management—The Genval Workshop Report. Gut 1999;44(suppl 2):S1–16.
4. Fass R, Fennerty MB, Ofman JJ, et al. The clinical and economic value of a short course of omeprazole in patients with noncardiac chest pain. Gastroenterology 1998;115:42–9.
5. Schenk BE, Kuipers EJ, Klinkenberg-Knol EC, et al. Omeprazole as a diagnostic tool in gastroesophageal reflux disease. Am J Gastroenterol 1997;92:1997–2000.
6. Shaw MJ, Talley NJ, Adlis SA, et al. Development of a digestive health status instrument. I: Tests of scaling assumptions, structure, and reliability in a primary care population. Aliment Pharmacol Ther 1998;12:1067–78.
7. Dimenäs E, Glise H, Hallerbäck B, et al. Quality of life in patients with upper gastrointestinal symptoms. An evaluation of treatment regimens? Scand J Gastroenterol 1993;28:681–7.
8. Mathias SD, Castell DO, Elkin EP, et al. Health-related quality of life of patients with acute erosive reflux esophagitis. Dig Dis Sci 1996;41:2123–9.
9. Young TL, Kirchodoerfer LJ, Osterhaus JT. A development and validation process for a disease-specific quality of life instrument. Drug Inf J 1996;30:185–93.
10. Locke GR, Talley NJ, Weaver AL, et al. A new questionnaire for gastroesophageal reflux disease. Mayo Clin Proc 1994;69: 539–47.
11. Carlsson R, Dent J, Bolling-Sternevald E, et al. The usefulness of a structured questionnaire in the assessment of symptomatic gastroesophageal reflux disease. Scand J Gastroenterol 1998; 33:1023–9.
12. Shaw MJ, Beebe T, Adlis SA, et al. Development of a digestive health status instrument. II. Assessment of validity, reliability, and responsiveness. Aliment Pharmacol Ther 2000 (in press).
13. Dillman DA. Mail and telephone surveys: The total design method. New York: John Wiley, 1978.
14. Jaeschke R, Singer J, Guyatt G. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.
15. Landis LM, Algina J. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
16. Stewart AL, Hays RD, Ware JE. Methods of constructing health status measures. In: Stewart AL, Ware JE, eds. Measuring functioning and well-being. Durham, NC: Duke University Press, 1992:67–85.
17. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
18. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill, 1994.
19. Talley NJ, Zinsmeister AR, Schleck AD, et al. Dyspepsia and dyspepsia subgroups: A population-based study. Gastroenterology 1992;102:1259–67.
20. Robins LN, Helzer JE, Croughan J, et al. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics and validity. Arch Gen Psychiatry 1981;38: 381–9.
21. Wittchen H. Reliability and validity studies of the WHO Composite International Diagnostic Interview: A critical review. J Psych Res 1994;28:57–84.